

Making sense of the world

The Earth and our effects on it require monitoring and analysis worthy of their complexity and importance. Now is the time to bring global observation into the twenty-first century.

Last week, ministers from some 60 nations gathered in Brussels to create an integrated Earth observation system, the Global Earth Observation System of Systems (GEOSS). December's tsunami in the Indian Ocean has catapulted GEOSS from relative obscurity to high on the international political agenda. This was clear from the presence of Carlos Gutierrez, the US commerce secretary, on his first overseas visit since being sworn in on 7 February, as well as science ministers from around the planet (see page 789).

The tsunami disaster highlighted the power of Earth observation data, but it has also thrown a harsh spotlight on the patchiness and rudimentary nature of current systems for understanding complex Earth systems and applying that knowledge to agriculture, management of water resources, early-warning systems for natural disasters, and more.

Take ocean currents, which affect climate by shifting large volumes of warm and cold water around the planet. A United Nations body set up in 1991 to observe, model and analyse the world's oceans, the Global Ocean Observing System (GOOS), has been chronically underfunded and has installed barely half of the monitoring instruments envisaged. Similar inadequacies undermine the Global Climate Observing System (GCOS) set up in 1992.

Speak to people working in global networks in almost any area of Earth observation and the message is the same: behind the stunning images and model simulations of planet Earth lies a much more disconnected picture. Countries and agencies tend to pursue their own agendas, resulting in duplication and a lack of sharing, and coverage is disproportionately concentrated in rich countries. Data come in overly diverse formats and units, making them difficult to use. Researchers, particularly those outside the Earth observation community, complain of costs, delays and other obstacles to getting the sorts of data they need.

Today's climate observation system is cobbled together from data from research satellites, weather satellites, atmospheric sounders and whatever ground-based observation stations scientists can get their hands on, rather than being tailored to monitoring and understanding climate change and variability. Research satellites last only a few years and are not replaced immediately, if at all. But for reliable climate-change monitoring over decades, it is essential, for example, to launch a follow-up satellite while its predecessor is still operating, so instruments on both can be cross-calibrated.

GEOSS is key to addressing such shortcomings. But researchers and other user communities should ensure that their needs are heard. Better international coordination promises to make better use of the billions of dollars that are already spent on Earth observation. But there is a limit to the benefits that can be squeezed out of coordinating and networking activities, when support for the basic scientific activity of collecting critical observations is neglected. Ultimately, GEOSS must make the case for, and oversee, an upgrading of systems such as GCOS and GOOS.

An optimistic view is that the political momentum to treat Earth observation more like global 'big science' facilities will translate into a better understanding and support of key scientific needs. The decision to house GEOSS within the World Meteorological Organization may bode well in this respect, as this Geneva-based UN agency has a good track record in mounting international operational weather systems.

But as the tsunami fades from memory, there is also a risk that the new-found political awareness will also subside. As GEOSS charts out what exactly it will do, scientists should actively engage with current political will by pressing home compelling arguments as to how better understanding of the spatial and temporal variability of the wide range of Earth-system parameters will result in progress. ■

Sooner than you think

Nature's back-page fiction is good for you.

What does the next half-century have in store? The record of the past fifty years shows that almost anything could happen. In 1955, the structure of DNA had been known only two years, and the complete sequence of the human genome wasn't even a distant prospect. Indeed, there were fewer than half as many humans as there are now. Roomfuls of vacuum-tube equipment were needed for computing power dwarfed by objects we now carry in our pockets. There were no cell-phones, no integrated circuits and almost no television. Antibiotics and transistors were novelties, the cold war a reality, and quarks existed only in James Joyce's *Finnegans Wake*. Space exploration was a dream yet to be realized.

In the same era, a generation inspired by the possibilities of science had taken an old 'westerns-in-space' formula and begun to forge a new kind of literature that asked serious questions about how technological change might affect the way we think about ourselves and other people. This was the golden age of science fiction. The 1950s

saw the publication of — to pick a few choice pebbles from the shore — Robert Heinlein's *The Man Who Sold the Moon*, Isaac Asimov's *Foundation*, Arthur C. Clarke's *Childhood's End*, Alfred Bester's *The Stars My Destination* and Walter Miller's *A Canticle for Leibowitz*.

In 1999 and 2000, *Nature* ran Futures, a series of science-fiction vignettes on what the coming millennium had to offer. Publishers know that, job-seekers apart, readers' attention wanes as they penetrate further into a magazine. So it may only be our most astute or compulsive enthusiasts who have noticed over the past few weeks that Futures has returned, on the back page of each issue.

Nature is proud to present Futures as a forum for the best new science-fiction writing, and the pieces — commissioned from well-established and novice writers alike — explore some of the themes that might challenge us in the next half-century or so. Prepare to be amused, stimulated, even outraged, but know this: the future is sooner than you think. ■





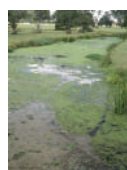
World view

Global monitoring system gets its feet off the ground
p789



In the clear

US drug agency unlikely to ban painkiller
p790



Fertile ground

Experts call for action over nitrogen pollution
p791



Dead loss

No way back from extinction for the Tasmanian tiger
p792

Tests in Tokyo reveal flaws in Vietnam's bird flu surveillance

David Cyranoski, Tokyo

Efforts to diagnose human cases of bird flu in Asia have been missing the mark, according to studies released last week. The findings have prompted calls for broader surveillance of the virus and a change in testing procedures.

Reanalysis of samples from Vietnamese patients with flu-like symptoms has revealed that some people originally declared free of bird flu actually did carry the avian virus H5N1. And another study has shown that some patients with symptoms other than those usually associated with flu were also suffering from H5N1.

The findings are worrying as they could mean that the virus has spread much more widely than was previously thought. This would, in theory, give the H5N1 strain a greater chance of mutating into a form that passes easily from person to person, potentially sparking a pandemic. Reassuringly, preliminary genetic analyses at the National Institute of Infectious Diseases (NIID) in Tokyo show that the virus has not mutated greatly since last year.

The recent findings could also lead experts to re-evaluate the severity of bird flu. The death rate for infected patients has so far been very high — 10 of the 11 cases identified in Vietnam since December 2004 have died. But if many more cases are going unidentified, the mortality rate could be much lower.

Samples from the 11 recent cases of bird flu in Vietnam, plus those from some 90 suspected cases that tested negative for H5N1, were recently sent to the NIID for study. About a third of the samples have been examined so far and, of these, seven of the negative results have tested positive, says Phan Van Tu, head of the microbiology and immunology department at the Pasteur Institute in Ho Chi Minh City, Vietnam. Fresh tests in Vietnam last week confirmed four of these positive results.

Part of the discrepancy between labs



Diagnosis of bird flu in Vietnam has missed some human cases.

could be accounted for by problems with the original tests. "Some reagents were not mixed well and the results weren't clear," says Tu.

But for three cases the Pasteur researchers — with NIID researchers observing — reconfirmed their earlier negative diagnosis. This is worrying as it suggests that the Vietnamese test is not sensitive enough to detect all cases. Tu says that the institute now plans to switch to the more sensitive test used in Tokyo and will ensure better training for its technicians.

If many infections have been missed, "suspected cases of human-to-human transmission should be investigated again," says a member of a World Health Organization (WHO) collaborating laboratory. There has so far been only one documented case of

probable human-to-human transmission (K. Ungchusak *et al.* *N. Engl. J. Med.* **352**, 333–340; 2005), but there are strong suspicions that clusters of disease within Asian families may have been transmitted through people rather than birds.

Concerns about H5N1 monitoring efforts have also been fuelled by a study showing that at least one case of encephalitis — a swelling of the brain — in Vietnam was caused by H5N1 (M. D. de Jong *et al.* *N. Engl. J. Med.* **352**, 686–691; 2005).

Previously it was thought that the only clinical symptom of H5N1 was respiratory disease. As encephalitis is relatively common in Vietnam, this could mean that a broader approach to H5N1 monitoring is needed, says Menno de Jong, a virologist at the Hospital for Tropical Diseases in Ho Chi Minh City, who led the study. The known cases of H5N1 may be just the "tip of the iceberg," he says. Although he notes that in an analysis of a further 100 encephalitis patients, none was infected with bird flu.

Other cases could have slipped through the net if the patients had mild respiratory illness, de Jong says, as only severe cases have been studied. He and his group have now begun a one-year survey to look for H5N1 among 1,600 children in Vietnam with mild symptoms.

In the meantime, many remain critical of current monitoring efforts. Henry Niman is the founder of Recombinomics, a company in Pittsburgh, Pennsylvania, that traces the molecular evolution of infectious agents. He claims that many cases of possible human-to-human transmission are not followed up properly. And a lack of widespread testing means that H5N1 could be to blame for other outbreaks, Niman says, such as a bout of apparent meningococcal blood poisoning that is ongoing in the Philippines.

There are no plans to test for H5N1 in the Philippines as yet. But the WHO is discussing how to broaden its surveillance of the virus. ■

REUTERS/CORBIS

Flu gene discovery prompts calls for tighter monitoring

Erika Check

Genes from a flu strain created in a lab in 1940 have been found in samples taken from pigs in South Korea, a US biologist claims.

Data from the flu virus samples were put last October in GenBank, the public database of genetic sequence information, by researchers at Chungnam National University in Daejeon, South Korea.

In December, the biologist Henry Niman of Recombinomics, a biotechnology company in Pittsburgh, Pennsylvania, examined the data as part of an analysis of flu sequences. He concluded that the samples contained genes from a strain of human flu virus that was created decades ago by scientists experimenting with the virus that caused the global flu pandemic of 1918.

Neither the World Health Organization (WHO), which coordinates the international response to flu, nor the South Korean government have commented on the claim. But Laurie Garrett, a former journalist and analyst at the Council on Foreign Relations in New York, says that the WHO attributes the sequence to an error at the lab that deposited the information.

Sang Heui Seo, one of the Korean researchers, says he is unable to comment yet, adding that "further confirmation" of the sequence "is under way at this moment".

The incident raises worrying questions about how the human flu genes got into a virus in a pig, says Niman. It could have happened naturally or in a lab accident — but it could also have resulted from experiments to produce a more deadly flu. "It could be bioterrorism," he says.

Niman says that the flu sequences posted on GenBank contain genes from a human strain called WSN/33, which was created in 1940 in a London lab. The strain was derived from a 1933 virus related to the one that caused the 1918 pandemic, and most people's immune systems have never been exposed to anything like it. The strain could be devastating if it infected people today, he adds.

Garrett thinks that even if there was a lab error, the episode should still ring alarm bells. Without a robust, global disease-surveillance network, she says, it is impossible to confirm the WHO's view that the sequence is a mistake. The discovery of the mystery strain "reveals critical weaknesses in our global security system", she contends. ■

New York draws fire over case of drug-resistant HIV

Erika Check, Washington

A decision by New York health officials to announce the detection of an unusually aggressive case of AIDS has led to criticism from some researchers and activists.

In December last year, a man from New York City tested positive for HIV and quickly showed signs of AIDS. Doctors at New York's Department of Health and Mental Hygiene believe that he developed AIDS between 2 and 20 months after he was infected; the disease usually takes about a decade to develop.

The patient's virus also resists treatment by the three important classes of HIV-fighting drugs. Officials at the health department say that this multiple drug resistance and the rapid progression to AIDS led them to warn the public of the possible spread of the strain. "This case is a wake-up call," the city's health commissioner Thomas Frieden told a press conference on 11 February.

Some scientists and doctors have praised the decision to publicize the case. They hope that it will warn the public about the dangers of having sex while under the influence of the drug crystal methamphetamine, or crystal meth. The infected patient had unprotected sex with many men while using the drug. "My hope is that this news will bring the reality to the public, and we will see less risky behaviour," says Jay Levy, a virologist at the University of California, San Francisco.

But others questioned whether the information released about the virus justified the public announcement. Similar cases have been reported before, critics say, but have not led to epidemics. In 2003, for example, researchers in British Columbia, Canada, reported two cases of HIV that seem similar to the New York virus. The Canadian patients were also infected with multidrug-resistant HIV that rapidly progressed to AIDS (K. C. W. Chan *et al.* *AIDS* 17, 1256–1258; 2003). But these viruses did not cause a widespread epidemic of HIV 'superstrains'.

"I don't think the health department needed to hang its public-health campaign on a single anecdotal virus," says John Moore, a virologist at Weill Medical College at Cornell University in Ithaca, New York.

The virus is being studied by researchers at the Aaron Diamond AIDS Research Center at Rockefeller University in New York. More data on the strain are scheduled to be released at the 12th Conference on Retroviruses and Opportunistic Infections in Boston this week.

But the Aaron Diamond's involvement has also drawn criticism. Some have even suggested that the centre pushed for the announcement to build interest in the retro-



Thomas Frieden is facing criticism for revealing the discovery of a virulent case of HIV.

virus conference, whose programme committee is chaired by David Ho, the Aaron Diamond's scientific director.

"There's a lot of suspicion because there's a confluence of issues, including the fact that the conference is around the corner, and David Ho is its chair," says Richard Jefferys, basic-science project director at the New York-based Treatment Action Group.

But Ho says that the health department made the call to alert the public about the virus after consulting with the US Centers for Disease Control and Prevention in Atlanta, Georgia. Ho adds that his group is presenting its data in the first available scientifically appropriate forum — the retrovirus conference. "I'm saddened by people who are trying to turn this into a personal attack rather than keeping focused on the case and its public-health ramifications," Ho says.

The health department also defended its action. "We had the necessary information and we were confident — and remain confident — that the situation was of great public-health significance," says an official. ■

Global observation project gets green light

Declan Butler, Brussels

Ministers from some 60 nations have signed up to a ten-year plan to build a unified, global Earth observation system.

The Global Earth Observation System of Systems (GEOSS) will work with existing agencies to distribute the data needed to address issues ranging from disaster mitigation and climate change to the management of water resources. It received formal backing at a meeting in Brussels on 16 February

The nations agreed to set up the system according to a ten-year plan developed by an ad hoc working group that was led by the United States, the European Commission, Japan and South Africa. The initial costs of the project, which will be agreed over the next two years, are estimated to be of the order of just tens of millions of dollars annually.

Last December's tidal waves in the Indian Ocean lent fresh impetus to the GEOSS proposal, its backers say. "The tsunami disaster has shown us just how important Earth observation can be, in providing data to support an immediate humanitarian response and subsequent reconstruction," says Janez Potočnik, the European Union's research commissioner.

If successfully implemented, GEOSS will coordinate the plethora of national and international agencies that already spend billions of dollars on Earth observation. And it will aim to provide more complete geographical coverage, on land, in the oceans, air and space, and over time.

Under the Brussels deal, a secretariat will be established at the World Meteorological Organization in Geneva, Switzerland. It will determine international

formats for data and build agreements under which agencies share data for free, or exchange it at low cost. Ultimately, GEOSS is also intended to broker funding and international agreements for new Earth observation instruments and facilities.

The Brussels meeting was the first overseas outing for the new US commerce secretary, Carlos Gutierrez. He enthusiastically backed the project, saying it would give us "the pulse of the entire globe".

Some scientists have been worried that GEOSS, which was first proposed in 2003, would serve mainly as a discussion forum, and a bureaucratic one at that. But at the meeting, they said that the scientific potential of the project was growing. "Potentially, GEOSS could grow into the most significant initiative in Earth observation since the invention of satellites," says Heiko Balzter, head of Earth observation at the UK-based Climate and

Land-Surface Systems Interaction Centre.

The project's most immediate boon may be data that are free and easy to use. Data-access policies vary widely between countries and organizations. "The costs quickly add up," says David Rogers, an ecologist at the University of Oxford, who works on satellite forecasting of malaria. Data are also often available only after a delay, or in user-unfriendly formats.

And such data can often be analysed only by experts, according to Paul Mason, former chief scientist at the UK Met Office and chair of the Global Climate Observing System (GCOS). "In cross-cutting programmes such as climate change, which require inputs from the land, ocean and atmospheric systems, it can be very difficult for individual research groups to use the data," he says, adding: "I hope GEOSS can change that."

Scientists also hope that the project will move Earth observation from a research activity to a more 'operational' footing, with services funded by governments worldwide, as for weather forecasting. "Basically, we need an operational system for climate observations," says Brian Hoskins, a climate modeller at the University of Reading, UK, and vice-chairman of the scientific committee for the World Climate Research Programme.

The GCOS programme, which reports to the United Nations Framework Convention on Climate Change, has been developing plans for such a system. "I welcome this GEOSS big brother, which will have the political muscle to pursue some of the difficult things that need to be done," says Mason. ■



Think global: GEOSS will coordinate Earth observations.

Russian security arrests institute head for spying

Bryon MacWilliams, Moscow

A senior Russian material scientist has been accused of selling state secrets to South Korea, in the latest of a string of arrests by the country's security services.

Oskar Kaibyshev, founder and director of the Institute for Metals Superplasticity Problems in Ufa, could be imprisoned for ten years if he is convicted. He says he is being accused of exporting dual-use technologies, which have both civilian and military applications, to a tyre manufacturer.

The 66-year-old researcher reported his own arrest on 18 February, telling news organizations that he had been questioned for some six hours on the previous day by agents of the Federal Security Service (FSB). He says that he has been suspended from his post at the institute, his bank accounts have

been frozen, and he has been prohibited from leaving Ufa, which lies 1,500 km southeast of Moscow.

The interrogation focused on several years of collaboration between the institute and the tyre maker ASA, a subsidiary of Hankook Tire, which is based in Seoul. Kaibyshev says that the firm is using superplastic technology in its designs for high-pressure tyres. The technology stretches titanium alloy to enhance its mechanical properties and, according to Kaibyshev, can be used to produce spherical tanks that can be inflated to a pressure of 1,000 atmospheres.

The institute has not had access to state secrets for two decades, Kaibyshev told the Moscow-based radio station, Ekho Moskvy. "If you need to put somebody on trial, it

should be agents of the FSB, who were fully informed of our contracts and should have warned us if there were any problems."

Ernst Chyorny, a member of the Public Committee for the Protection of Scientists — a Moscow-based human-rights group — says that the technologies in question had already been exported to India in 1987, and to Italy in 1990. They are also described in a forthcoming book co-authored by Kaibyshev. The book is being funded by the International Science and Technology Center, a Moscow-based project supported by, among others, the United States and the European Union. The centre helps military scientists from the former Soviet Union to find civilian work.

The FSB declined to comment, saying that it would release details of the charges against Kaibyshev next week. ■

Speech transcript stokes opposition to Harvard head

Emily Singer, Boston

The president of Harvard University is under growing pressure, following the release of a full transcript of contentious remarks he made last month about women in science.

Some faculty were considering calling for a vote of no confidence in Larry Summers at a special meeting of the Faculty of Arts and Sciences due to take place on 22 February, although there are indications that such a vote may not take place until next month.

National controversy continues to rage over Summers' comments, in which he suggested that differences in "intrinsic aptitude" might be a key factor behind the scarcity of women in science, outweighing the impact of gender discrimination.

The Harvard Corporation, the board to which Summers is answerable, took the highly unusual step of issuing a statement of support for him. "We are confident of his ability to work constructively with the faculty and others," the 17 February statement said.

The same day, Harvard released a 7,000-word transcript of Summers' speech and the questions that followed it. In the transcript, Summers compared the under-representation of women in science to that of whites in the National Basketball League and Jews in farming.

Summers' position was damaged by a stormy, closed-faculty meeting on 15 February, at which many criticized what they see as his autocratic management style (see *Nature* 433, 190–192; 2005).

Some critics say that the release of the transcript has weakened his position. "My sense is that Larry Summers' job is on the line," says Everett Mendelsohn, a historian of science at Harvard.

"He has only a few options," says Arthur Kleinman, chairman of the anthropology department. "Resign, get fired, or rework his policy style." Fellow economists, however, have circulated a letter of support for Summers, which by 21 February had attracted 180 signatures.

And even some critics of his remarks on women say he should stay. In her research, Harvard psychologist Elizabeth Spelke has found no evidence of a biological basis for gender differences in mathematical ability. "I disagree with almost everything Summers said," says Spelke. "But I hope he will keep his job and rectify the situation, in the way he said he would." Summers has repeatedly apologized for his remarks. ■



Pill for all ills: painkillers such as Vioxx increase patients' risk of stroke but are unlikely to be banned.

Vioxx may go back on sale after scraping past FDA panel

Meredith Wadman, Washington

Advisers to the US Food and Drug Administration (FDA) have voted, by a narrow margin, that it should not ban Vioxx — the painkiller withdrawn by drug-maker Merck.

They also said that Pfizer's Celebrex and Bextra, two other members of the family of painkillers known as COX-2 inhibitors, should remain available, despite the fact that they too boost patients' risk of heart attack and stroke.

The recommendations of the arthritis and drug safety advisory panel offer some measure of relief to the pharmaceutical industry, which has faced a barrage of criticism for its promotion of the painkillers. But the advice of the panel, which met near Washington DC over 16–18 February, comes with several strings attached.

For example, most panel members said that manufacturers should be required to add a prominent warning about the drugs' risks to their labels; to stop direct-to-consumer advertising of the drugs; and to include detailed, written risk information with each prescription. The panel also unanimously stated that all three painkillers "significantly increase the risk of cardiovascular events".

The panel voted 17 to 15 against banning Vioxx (rofecoxib) entirely; the vote on Bextra (valdecoxib) was 17 to 13 with 2 abstentions; Celebrex (celecoxib) was endorsed 31 to 1. Shares of Merck, based in Whitehouse Station, New Jersey, and New York-based Pfizer closed up 13% and 7% respectively on 18 February, the day of the votes.

The FDA is expected to act on the recommendations within weeks. Although the agency usually follows the recommendations of its outside advisers, it is not bound to do so. A top official said that, in light of the closeness of some of the votes, the agency will

examine the panel members' comments in detail before deciding what to do.

An official from Merck said during the meeting that it would consider reintroducing Vioxx, which it withdrew in September 2004. Pfizer's two painkillers are still on the market.

Throughout the three-day meeting, panelists struggled to reach a decision that balanced benefits with undeniable risks. They complained of a dearth of comprehensive clinical data on the cardiovascular risks of the drugs, and a similar lack for conventional non-steroidal anti-inflammatory inhibitors. Many physicians will now fall back on the latter for treating arthritis, despite the fact that they can cause gastric bleeding.

Alastair Wood of Vanderbilt University Medical Center in Nashville, Tennessee, the panel chairman, noted that the committee was dealing with "a much bigger safety problem" than existed for any of the drugs the FDA has withdrawn in the past. Robert Temple, acting director of an FDA Office of Drug Evaluation, said that the kind of study needed to address safety questions on the painkillers "is mind-bogglingly large", requiring perhaps 50,000 subjects.

The meeting capped a week of important developments at the FDA. On 14 February, President George W. Bush nominated Lester Crawford, acting head of the agency, as its permanent commissioner. The next day, the agency said it would create an advisory board to monitor and publicize safety problems in drugs already on the market.

But this Drug Safety Oversight Board will be dominated by FDA officials and will lack the power to ban drugs. "It's a cosmetic gesture," says David Graham, the physician in the agency's Office of Drug Safety whose high-profile Senate testimony last November first brought the agency under intense scrutiny. ■

Nitrogen study fertilizes fears of pollution

Jim Giles, London

Urgent political and scientific action is needed to tackle the global threat of nitrogen pollution, say scientists behind one of the field's biggest research projects.

They gathered in London last week to mark the completion of a five-year, £7-million (US\$13-million) project to map the effects of excess nitrogen on forests, rivers and grasslands, primarily in Britain.

The researchers say that the Global Nitrogen Enrichment (GANE) programme has transformed their understanding of how nitrogen affects the environment. But a lack of similar studies in other countries has led to inadequate legislation, which is generating a growing threat to global biodiversity, they say. Previous efforts have been made to draw attention to the issue, but researchers say that much more needs to be done.

Since the Industrial Revolution, humans have been converting unreactive nitrogen gas from the atmosphere into reactive forms such as ammonia, primarily for use in fertilizers (see graph). These industrial processes have revolutionized food production, but, along with the release of other reactive forms of nitrogen from the burning of fossil fuels, they have led to a massive increase in the amount of reactive nitrogen in circulation.

"This is the third major threat to our planet after biodiversity loss and climate change," says John Lawton, chief executive of the Natural Environment Research Council, which provided the bulk of the funding for GANE. "It's manifestly unsustainable in the long term."

Excess nitrogen is known to have a variety of ill effects on plant life. Parts of the Gulf of Mexico, for example, are so inundated with excess fertilizer that the water is clogged with algae, suffocating fish and other marine life.

GANE researchers wondered whether a detrimental effect is also occurring in freshwater lakes. They surveyed 60 shallow freshwater lakes in Britain and Poland to see whether dissolved nitrate ions are involved in observed losses of plant life. Brian Moss, an ecologist at the University of Liverpool, who led the study, says the result is clear: "More nitrate means fewer species."

Nitrate ions tend to boost the growth of one or two plant species, says Moss, which go on to dominate the lake. If these dominant species are lost, perhaps during a cold snap, algae can take over and prevent any plant life from returning. Phosphorus is known to have this effect, and the GANE results suggest that nitrogen is equally important.

Moss fears that European regulations on nitrates are too lax to prevent further plant loss. In Britain, for example, government regulations stipulate that nitrogen in nitrate ions in water must not exceed 10.3 milligrams



Spreading menace: nitrogen-based fertilizers enter rivers and cause them to become choked by algae.

per litre, primarily for reasons to do with human health. But according to Moss, levels of 2–3 milligrams per litre can wipe out all but a few species from most shallow lakes.

The results will not be published until later this year, but Moss says he made the UK government aware of the finding two years ago. A spokesman for the Department for Environment, Food and Rural Affairs says that nitrogen research is a "developing area of

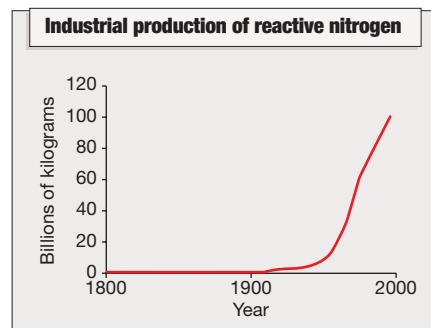
Carolina is one such area, says Alan Davison, an expert in air pollution at the University of Newcastle upon Tyne, UK, and scientific coordinator for GANE. "There are 120 species of tree there — equivalent to the whole of Europe," he says. Nitrogen levels in the area are known to be high, he adds, and there is evidence from other forests that this could be detrimental (see *Nature* 425, 894–895; 2003).

Other biodiversity hotspots may also be threatened. Europe and North America have previously been the biggest sources of reactive nitrogen, but by 2020 half of anthropogenic nitrogen is expected to come from developing nations, which are home to many of the world's most species-rich areas.

Nitrogen pollution has not been properly measured in many of these regions. An attempt to estimate the nitrogen load on biodiversity hotspots has been completed, although the research team, led by University of Sheffield plant biologist Gareth Phoenix, has declined to reveal details until its paper has been peer reviewed.

Other groups are also trying to raise the issue's political profile. The International Nitrogen Initiative promotes cooperation on nitrogen research, and last year produced a review summarizing the predicted impact of global nitrogen pollution.

The group's chair, biogeochemist James Galloway of the University of Virginia in Charlottesville, is also one of the researchers behind the Nanjing Declaration on nitrogen management. This document, which was signed by scientists last October in Nanjing, China, calls on the United Nations to promote nitrogen research and to consider policy responses.



Global production of nitrogen-based fertilizers and explosives has been rising dramatically.

science" and that government policy cannot be revised until more is known about the impact of nitrogen on freshwater ecology.

The conference also heard evidence that nitrogen levels are linked with sometimes subtle changes in grassland species, from mosses and lichens to grasses and flowers. GANE scientists urged the government to take account of these results, but lamented that similar data are not available for other parts of the world that face a similar threat.

The forests of the Great Smoky Mountains on the border of Tennessee and North

Survey reveals extent of tsunami havoc along Sri Lankan coast

San Francisco Researchers on a geological survey of the tsunami-hit Sri Lankan coast have released their preliminary results.

The International Tsunami Survey Team, which flew out a fortnight after the waves struck on 26 December, collected data and eyewitness reports for some of the worst-hit coastlines in the east and south of the island (see *Nature* 433, 350–354; 2005). The team observed that even neighbouring beaches showed very different levels of flooding and destruction, depending on coastal orientation and underwater topography. The limit of the tsunami's reach also varied widely — in some places it flooded inland areas more than 12 metres above sea level.

The US Geological Survey team says the information can be used to plan rebuilding in the aftermath of the disaster, and to calibrate models of tsunami behaviour against field observations.

► <http://walrus.wr.usgs.gov/tsunami>

Museum drops plans to clone extinct marsupial

Sydney The Tasmanian tiger is to remain extinct — for now, at least.

Plans to resurrect the thylacine (*Thylacinus cynocephalus*), a dog-like marsupial, using DNA samples from a preserved specimen were abandoned by the Australian Museum in Sydney on 16 February. The museum says the samples are too degraded to be used, but



Lost stripes: preserved DNA samples of the Tasmanian tiger are too degraded for cloning.

advocates of the project say they will explore other options.

The last-known thylacine died in captivity in 1936. But in 1999, the museum's director Michael Archer championed an initiative to clone the species using a related marsupial, the Tasmanian devil, as a surrogate mother. The proposal generated considerable scepticism in scientific circles.

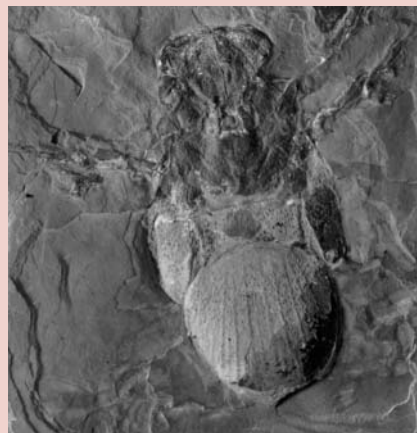
Archer, who became dean of science at the University of New South Wales in Sydney

Biggest-ever spider recast as a sea scorpion

London A giant 'tarantula' that lived some 300 million years ago is not, after all, the largest spider that ever walked the Earth. It is a sea scorpion, scientists reported last week.

The fossil of *Megarachne servinei* (pictured), which had a half-metre leg span, was discovered in 1970 in Argentina by Ernesto Servine, an amateur collector; it was classified as a spider by Argentinian palaeontologist Mario Hünicken. The creature has an entry in the *Guinness World Records*, but doubts have long existed about the classification.

Servine kept the fossil in a bank vault, and scientists only had the opportunity to re-examine it after he died. Paul Selden, a palaeontologist at the University of Manchester, UK, has now declared it to be a sea scorpion after comparing it last year with a newly discovered specimen from the same locality



and with *Woodwardopterus*, a sea-scorpion species found in Scotland.

"It took a long time to get to see it — now we're happy," said Selden.

last year, says he is disappointed by the decision. "But I and other colleagues remain interested, and I don't think the project will die because the museum can't proceed," he adds.

Women lack support, not drive, study finds

London A survey of scientists has cast doubt on the notion that female scientists fail to reach top positions because they are less ambitious than men.

The study, which analysed results from more than 6,500 UK scientists, found that women in research institutes have higher career aspirations than men. But when it comes to applying for group leader and senior management positions, women get significantly less encouragement and support from tenured researchers and department heads, the authors said on 18 February.

The research was carried out by the Athena Project, a government-sponsored scheme designed to advance women's careers in science, and the University of East Anglia in Norwich.

"The sad fact is that those further down the scale are still not rising to more senior positions with the same frequency as men, despite a clear desire to do so," says Caroline Fox, Athena's programme manager.

► www.athenaproject.org.uk

Governor attacks state backing for stem-cell work

Washington Moves enabling the state of Massachusetts to become a supporter of stem-cell research are coming under attack.

The state's lawmakers are considering a proposal that would permit researchers to clone embryos in order to produce

stem cells. The development would add momentum to plans to conduct similar research at the new, privately funded Stem Cell Institute at Harvard University (see *Nature* 428, 8; 2004). But these plans were thrown into doubt on 10 February, when state governor Mitt Romney said he opposes such practices and would like to ban them.

Romney's statements have prompted angry criticism from some quarters. Observers suggest that his real motivation is to position himself favourably with the national Republican Party, which opposes the creation of embryos for research.

Super sub is frozen out after failing to return

London A pioneering robot submarine has become trapped in a dark and cold ocean grave — under an Antarctic ice sheet.

Autosub, operated by the Southampton Oceanography Centre, is a veteran of almost 400 missions spanning the past four years. Earlier this month, it set a landmark for marine exploration by journeying under the permanent Antarctic ice sheet to investigate rising sea levels (see *Nature* 430, 955; 2004). But the vessel has failed to return from a repeat mission that began last week. Its owners said on 18 February that the sub is trapped 17 kilometres from the edge of an ice shelf and probably cannot be recovered.

The 7-metre-long robot, powered by some 5,000 household batteries, is programmed to measure the thickness of the ice sheet and to collect water samples. A replacement is under construction and should be ready by September.



D. TEMPLETON/ALAMY

FLASH IN THE PAN?

Obesity is the main target in the US government's latest dietary guidelines. But can this advice really make a difference? *Nature's* reporters sift through the heady mix of politics and science to get a taste of things to come.

"If you follow this diet, you're going to lose weight, you're going to be healthy and you're going to be able to improve your quality of life ... it's scientifically based, but it's also common sense."

Another diet guru flogging their snake-oil prescription for the servings of fat, carbohydrate, protein and other nutrients needed to be healthy and slim? No. This was Tommy Thompson, then US Secretary of Health, speaking on 12 January at the release of Uncle Sam's very own diet book, *Dietary Guidelines for Americans 2005*.

The guidelines, revised every five years, inevitably amount to a compromise between nutrition advocates and the food and agriculture lobbies. Yet this time they largely have pleased even staunch critics of government food policy. "They look to me like the strongest dietary guidelines yet produced," said Michael Jacobson, who heads the Center for Science in the Public Interest, a nutrition advocacy group in Washington DC, at a press conference after the release.

But in the aftermath, a philosophical divide has emerged. On the one hand is the view, expressed by Thompson, that the government's role is to put out information about what constitutes healthy eating, but that it's up to individuals whether they follow the advice. The other take is that the government must do more, not only to educate people about food choice, but to ensure a food supply that accurately reflects its own dietary advice. For those who take this view, the guidelines don't go far

enough — and they say that buried in the fine print are concessions to the food industry that threaten to weaken the impact of the advice.

The guidelines — a joint effort by the Department of Health and Human Services (DHHS) and US Department of Agriculture (USDA) — were first issued in 1980, and form a reference for US eating habits. They underpin government nutritional policy and federal food programmes, including school meals. And they will be summarized graphically in a new 'food guidance system', which will be released within weeks to replace the 'food pyramid' introduced in 1992.

Fat fighters

Previous guidelines focused on cutting consumption of the saturated fats that cause chronic conditions such as heart disease. But the top priority now is to roll back the obesity epidemic that is causing a surge in conditions such as type-2 diabetes. About two-thirds of US adults are deemed overweight or obese by the Centers for Disease Control and Prevention in Atlanta, Georgia.

Consequently, the biggest change in the new guidelines is the emphasis on restoring energy balance to people's diets. The message is that there is no getting round the laws of thermodynamics. If your calorie intake exceeds your energy output, you will gain weight. To this end, the guidelines advise a close watch on calories and 30–60 minutes of exercise most days of the week; 90 minutes to shed unwanted flab.

C. SMITH/HHS

That's old hat to dieters, but the surprise this time was the explicit statement that the healthiest way to reduce calories is to avoid added sugars, certain fats and alcohol, all of which are high in calories but low in essential nutrients. In the past, such a message has been all but taboo. The previous guidelines say only that added sugars may contribute to weight gain. And in January 2004, the Bush administration lobbied against phrasing in the World Health Organization's dietary advice that urged people to eat fewer sugary and other high-calorie foods.

To hammer its point home, the scientific advisory committee behind the guidelines turned to the concept of 'discretionary calories': the number of junk-food calories you can eat daily without gaining weight. A typical sedentary person who burns 2,200 calories per day needs to eat about 1,910 calories of healthy food to meet their basic nutritional needs. This leaves 290 calories for a treat, such as beer and potato crisps with late-night TV.

The idea, say committee members, is to raise people's consciousness about overeating without denying them their favourite snacks. Even a relatively modest reduction of between 50 and 300 calories per day could prevent most new cases of obesity, particularly among children. "If we could achieve this it would be the major crowning achievement of the guidelines," says Alice Lichtenstein, a cardiovascular researcher at Tufts University School of Medicine in Boston, Massachusetts, who sat on the scientific committee for the 2000 guidelines.

Taste of the future

Also new this year is an emphasis on fibre- and nutrient-rich whole grains instead of refined grains, and a recommendation to eat nearly twice the quantity of fruit and vegetables suggested in the 2000 version, as a way to lower the risk of certain cancers, type-2 diabetes, stroke and obesity. And the guidelines now clearly distinguish between different types of fat. Gone is the blanket low-fat creed of the past 20 years. In its place is advice to avoid saturated fats, which are found in red meat, for example, and *trans*-fats, which are abundant in processed foods. At the same time, moderate amounts of healthy fats, such as olive oil, are recommended.

The stronger wording in the new guidelines is partly the result of changes to the drafting procedure that gave scientific advisers greater autonomy. In 2003, the DHHS and USDA appointed 13 nutrition scientists to the Dietary Guidelines Advisory Committee and asked them to compile a report from the latest scientific literature. Bureaucrats at



The US government's guidelines for a healthy diet were released in January by Tommy Thompson (above), but some say influence from parties, such as the sugar industry, will compromise the advice.



the two agencies then reviewed this report, published in the *Federal Register* in August 2004, and considered comments from interested parties. They drafted the guidelines themselves, and communications specialists transformed them into the slick, 84-page brochures released in January.

Separating the two phases made the scientific basis of the guidelines more transparent, says Janet King, chair of the committee and a researcher at the Children's Hospital Oakland Research Institute in Oakland, California. In the past, the committee members had to write the actual guidelines, which forced them to consider factors such as how easy they were for a lay reader to understand. The previous committee spent ages, for example, debating whether to advise people to 'limit' or 'moderate' their salt intake, King says. The new set-up also made the committee less of a target for pressure from industry and policy-makers. "It shielded us," she says.

But the literature is far from definitive about the best diet. There are few long-term or well-controlled clinical nutrition trials available, so the committee relied heavily on epidemiological and observational studies (see

'How is science converted to dietary advice?' page 798 and *Nature* **428**, 252–254; 2004).

These uncertainties left plenty of room for quibbling about the wording of the guidelines, say critics of the process, who have been quick to point out the fingerprints of the food industry in the small print. One charge levelled at this and previous guidelines is that they tell consumers only what foods they ought to eat — such as lean cuts of meat and low-fat dairy products — without spelling out foods to avoid, such as processed snacks, fast food or red meat dripping with saturated fat. "They're saying the right thing but not quite giving it the teeth it needs," says Carlos Camargo, himself a member of the scientific committee and an epidemiologist at the Harvard School of Public Health in Boston.

Cream of the crop

A more specific industry influence, critics contend, is the recommendation to consume more dairy products — the equivalent of three cups of milk a day, up from between two and three cups last time. They say that this increase provides most people with unnecessary calories, that it is possible to achieve recommended intakes of calcium and other nutrients through other means, and that it fails to take into account studies linking diets high in dairy products with an increased risk of prostate cancer. The increase is "one of the strongest influences of the food industry" in the report, says Walter Willett, a nutrition epidemiologist at the Harvard School of Public Health.

Concerns have also been raised about the

D. NUNUK/SPL

recommendation on *trans*-fats. The scientific report said that these should not exceed 1% of daily energy intake, but the guidelines say only that *trans*-fat intake should be kept “as low as possible”. According to Camargo, this was the most significant departure from the committee’s recommendations, allowing those putting together school meals, for example, to make only a token effort to reduce the fat. Nutrition researchers critical of the food industry charge that the language was softened under industry pressure to avoid costly revamping of production processes that rely on cheap vegetable oils.

For its part, the food industry says it is just trying to keep the guidelines fair. Industry representatives, as well as other interest groups, were invited to provide the committee with written and oral comments during the literature-review phase and to make comments after its report was published. The National Dairy Council, for instance, presented evidence supporting its argument that dairy foods help people meet their calcium and potassium requirements. And the Grocery Manufacturers of America argued that added sugars help increase the palatability of some nutritionally valuable foods.

Reaping benefits

But industry has other avenues of influence open to it. One is through USDA, which, by the nature of its mission, is more attuned to farmers’ interests than to public-health needs. “It’s the wrong agency to do this, and a blatant conflict of interest,” says Marion Nestle, a prominent critic of the food industry working at New York University.

A second and more opaque route is through lobbying — an integral part of the US political system — where industry and others try to influence agency officials by, for example, providing them with relevant documents and making personal contacts.

Thompson openly discussed industry’s influence at the launch of the guidelines. “The food industry has spent a great deal of time and money appearing in and observing all of the negotiations and all of the testimonies that went into compiling the guidelines,” he said. “They come in and meet with me on a regular basis.”

Although industry may have won key concessions, anyone who follows the guidelines strictly will probably end up in better health. The reason some nutrition experts are still not happy is that they anticipate little time or money will be put into spreading or enforcing the advice.

“What is lacking is will on the part of the government and Congress to convert the guidelines into new health and agriculture policies and programmes,” says Jacobson. He asserts that doing so would step on major interests such as restaurants, as well as the corn, sugar, processed-food and salt industries. What is needed, he says, are hard-



Eat yourself fitter: the new US guidelines encourage greater consumption of fruit and vegetables.

hitting, mass-media campaigns to help shift consumer demand to healthier products. Some also advocate legislation to subsidize healthy foods, regulate advertising aimed at children, and to require calorie information to be displayed on restaurant menus.

So far there is little sign that strong implementation is coming. Instead, the US administration has emphasized that diet is a matter of personal choice. “It’s up to the individual to make the right decisions,” said Thompson in January.

But Ricardo Uauy, an expert on health and nutrition at the London School of Hygiene and Tropical Medicine, argues that it is nearly impossible to choose carbohydrates or fats “wisely”, as the guidelines recommend, when many children’s cereals contain as much as 40% sugar, for instance, and processed foods account for 80% of the *trans*-fats that people eat. Nor is it easy for people to choose foods with little salt — as

the guidelines advise to lower blood pressure — when 80% of their salt intake comes from processed foods.

Hard to swallow

Indeed, the US diet will have to change radically to meet the new advice. At present, many Americans eat a diet that resembles the food pyramid turned upside down, with too much salt and added sugars and fats, and not enough grains, fruit and vegetables. Without substantial changes in the practices of the food industry the guidelines will have little impact, predicts Uauy.

Others say that the food industry is

already going through a period of evolution spurred in part by consumer demand. The public discussion of obesity and related health problems has led to increased awareness among consumers and greater scrutiny of the industry by nutrition advocates and the media. Many fast-food chains have updated their image with healthier fare such as salads and yoghurts. Last year, McDonald’s phased out its ‘supersize’ meal options. And PepsiCo has removed *trans*-fats from some of its snacks and has introduced a greater range of bottled waters and fruit juices as alternatives to sugary drinks.

At the same time, food companies are not about to abandon their calorie-laden products as long as demand for them exists. “If people want French fries and a double cheeseburger we’re gonna give them that,” says Bob Goldin, executive vice-president of Technomic, a food-industry consulting and research firm based in Chicago.

But industry experts say that the shift towards healthier foods is more than cosmetic. Health and whole foods are one of the biggest growth areas in an otherwise saturated market, and companies are scrambling for a share of it. They predict that consumer demand for healthier food will grow, partly as a result of the dietary guidelines. “It will be a driving factor in the industry going forward, because that is what the consumer will ultimately want,” Goldin says.

Certainly Americans are hungry for food advice, if \$2 billion in diet book sales last year is any indication. But whether it will take more than a few books and a gentle nudge from their federal health department to get them to eat better and slim down is still up in the air.

Declan Butler and Helen Pearson

♦ www.healthierus.gov/dietaryguidelines

Around the world in three square meals

In China, it's a pagoda; in Canada, a rainbow. But despite the diversity of design to be found in pictorial food guides from around the world, the core advice remains the same: eat your peas and porridge, limit your bacon and eggs.

In Canada's rainbow, for instance, breads and cereals occupy the outermost — and therefore longest — curve of the arc. This band is coloured golden yellow to represent grain. Vegetables and fruit are next in green, followed by dairy in blue and meat in the diminutive, innermost red band.

Of course, in real rainbows, even Canadian ones, red comes before yellow. Putting the colours in that order comes a bit closer to what real Canadians eat. According to data from the United Nations Food and Agriculture Organization, Canadians, like other North Americans, get more than twice as many calories from meat and fish as from fruit and vegetables.

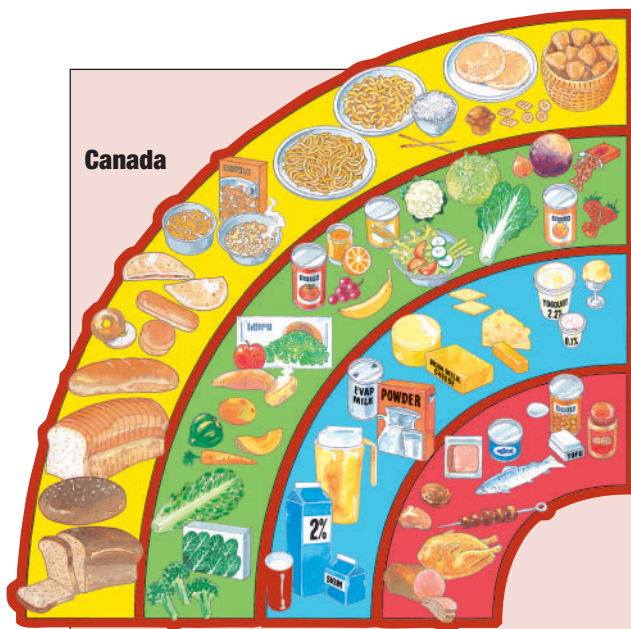
That ratio is fairly common in parts of the world where meat is readily available. The Chinese, avid consumers of pork, have a similar proportion in their diets. This is despite the advice of the Food Guide Pagoda, which has grain in its foundation level, and fruit and veg just above it.

Nearly every official food guide emphasizes grains and cereals as the foundation of a healthy diet, and that's one recommendation the world as a whole has no trouble living up to. In the United States and Europe, grains and cereals make up about a quarter of the average diet. In Asia, where rice is a staple, they are anywhere from 50% to 60% of daily calories. Diets in most other regions fall somewhere in the middle.

Although the basic recommendations are the same, each pyramid, rainbow or circle tends to reflect the nation's unique food culture. The Mexican food circle has an entire wedge devoted to beans. The Chinese pagoda's food depictions include a bowl of rice and a head of bok choy, and the German food circle features photographs of hearty whole-grain breads.

But no food guide seems to take adequate account of the irrepressible human sweet tooth. Sweets are listed along with fats as only for occasional consumption in most guides. And several make no mention of sweets at all, including those from China, Sweden, Germany and Portugal (see J. Painter, J.-H. Rah and Y.-K. Lee *J. Am. Diet. Assoc.* **102**, 483–489; 2002). Even so, North Africans get 9% of their calories from sugar, Europeans 11%, and Americans a cloying 18%.

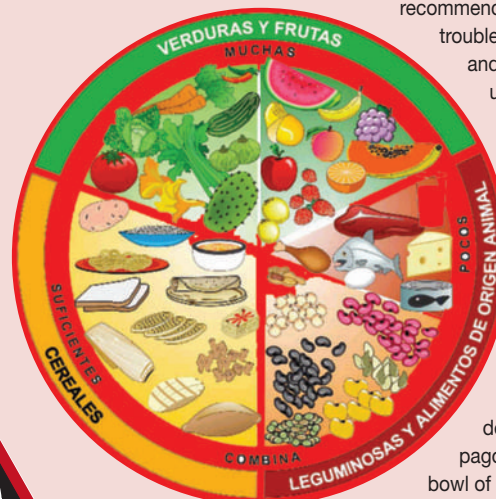
Jonathan Knight



China

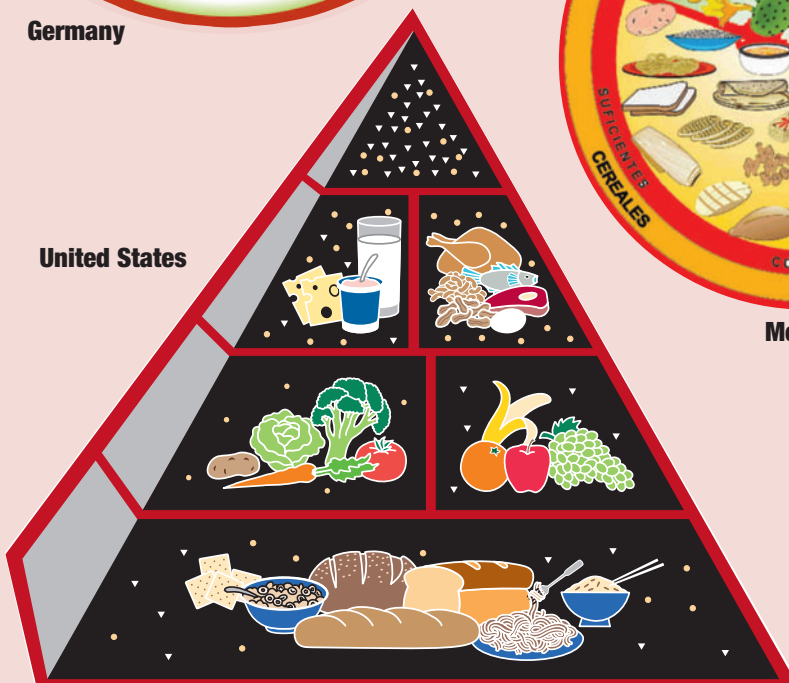


Germany



Mexico

United States



Food FAQs

Eating a healthy diet is hard work. There are hundreds of guides out there — often providing conflicting instructions. Deciding what advice to take means wrestling with a number of tough questions.

How is science converted to dietary advice?

To make the jump from scientific data to specific recommendations for a healthy diet — that everyone should now eat nine servings of fruit and vegetables a day, for instance — the scientists on the US Dietary Guidelines Advisory Committee turned to computer modelling.

They mixed and matched foods from the traditional groups — fruits, vegetables, grains, meat and beans, dairy foods, fats and sweets — to come up with combinations that met nutritional requirements put forward by the Institute of Medicine's Dietary Reference Intakes. This gave them a series of food patterns for a range of daily calorie intake levels from 1,000 to 3,200, in increments of 200. These model diets were designed to boost nutrients that are often too low in US diets, such as vitamin E, calcium, magnesium, potassium, fibre and vitamin A.

In a twist, the nutritional value of each food group was calculated to reflect what Americans actually eat, rather than as a simple average of all the foods in the group. For example, broccoli accounts for more than half of the greens Americans eat, and spinach about a fifth, according to the 1999–2002 National Health and Nutrition Examination Survey. So the value of the greens group as a whole was calculated as having 0.53 of the nutritional values of broccoli and 0.20 of those of spinach. The remainder was taken as the combined average of other greens.

Committee members say that this approach allowed them to make recommendations with practical value to Americans, who are not likely to radically shift the proportions

of foods they eat within the food groups.

But the result did not entirely satisfy nutrition professionals. Dena Bravata, a physician and obesity researcher at Stanford University in California, says that although the guidelines are valuable she would have preferred them to be “based more on the scientific evidence rather than this hybrid approach”. The average US diet is hardly ideal, she says, and knowing what food combinations are optimal would allow patients and clinicians to create individualized diets based on the best available evidence.

The approach has its disadvantages, the committee admits. Many people in the United States don't get enough vitamin E, for example, so one might have expected the guidelines to recommend eating more nuts and oils, which are rich in this vitamin. But Americans eat very few real nuts (peanuts, although popular, are actually legumes) and use oils that are low in vitamin E, so nuts and oils ended up with low vitamin E scores in the computation. This meant acrobatic accounting to boost other food groups with average levels of vitamin E. **Declan Butler**

Which countries or cultures have the best diets?

Several groups are in the running. Many people consider the traditional Mediterranean diet to be one of the healthiest. But times are changing. With the globalization

of the food market, processed foods are creeping into traditional diets at the same time that physical activity is declining in many parts of the world. More and more, the people with the best diets are those who make a concerted effort.

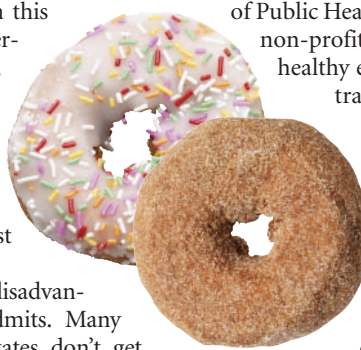
In collaboration with the Harvard School of Public Health, Oldways, a Boston-based non-profit organization that promotes healthy eating, has assembled several traditional diets into food-guide pyramids, following the shape of the official eating guide set out by the US Department of Agriculture. These take traditional dietary patterns into account, as well as data from clinical and epidemiological research.

The Oldways Mediterranean pyramid is based on the diet of Greece, Italy, Portugal and Spain around 1960, a time when people in those countries lived longer than their northern European neighbours and were less likely to develop heart disease. Their daily fare included wholegrain bread, pasta, rice, fruit, beans, vegetables, cheese, yoghurt and that quintessential Mediterranean ingredient, olive oil. They also ate fish, poultry, eggs and sweets weekly. But red meat, with its artery-clogging saturated fat, was consumed less often.

Today, these eating habits are gradually being abandoned, and at the same time the Mediterranean advantage in life expectancy has decreased¹ and obesity is on the rise². But



Fruitful analysis: researchers have modelled the contribution of various food groups to the US diet.



STILL PICS



the diet itself remains popular among those who try to eat right. As a result, diet-conscious people in places that lack strong nutritional traditions may be candidates for the best eaters today. "Middle-aged, educated women in California eat particularly healthy diets," says Martijn Katan, nutrition scientist at the Wageningen Centre for Food Sciences in the Netherlands.

If effort is essential, certainly few countries have tried harder to eat right than Finland. In the early 1970s, Finnish men had the highest rate of heart disease in the world. Their diet consisted of large amounts of whole milk, cheese and salt, with very little in the way of fruit and vegetables. A national programme of education and changes to the food supply over three decades has vastly improved the national diet. Today, the mortality from coronary heart disease in working-age men has been reduced to a quarter of what it was in the 1970s (ref. 3). **Achim Schneider**

There must be a natural diet for humans — what did we evolve to eat?

The good news is that evolution teaches us that humans can eat just about any mix of the basic food groups. During evolution, we have colonized almost every ecosystem on Earth, and adapted to what was available; from Arctic populations eating almost exclusively animal protein, to villagers in the peaks of the Andes living largely on grains and cereals.

We evolved as "flexible eaters," says William Leonard, an anthropologist at Northwestern University in Evanston, Illinois, and an expert on diet in evolution.

Taken from this evolutionary vantage point, arguments in diet books over whether a low-fat, high-carbohydrate diet is better than a high-protein, low-carb diet make no sense, he says. Alice Lichtenstein, a cardiovascular researcher at Tufts University School of Medicine in Boston who sat on the scientific committee that produced the 2000 US dietary guidelines, agrees. "A variety of diets are possible from an evolutionary point of view," she says.

The bad news is that evolution has also left us with a genetic legacy: our brains and genes are hardwired to seek out as much energy as possible for the least physical effort. This served humans well during millennia when starvation was a constant threat to our survival, but is not adapted to the modern world where high-calorie foods are a phone order or short drive away.

Evidence for this can be found in the Arctic, for instance. Indigenous people who maintain a traditional lifestyle eat a great deal of meat, yet they have low blood lipid levels, which is a risk factor in heart disease, and enjoy good cardiovascular health. The explanation, Leonard proposes, is that their rate of metabolism is raised as a result of vigorous physical activity and in response to their frozen environment. But their relatives who have adopted a more sedentary way of life, and a Western diet that has more processed foods and less meat, have significantly increased blood lipid levels and higher rates of obesity and cardiovascular disease⁴.

Such findings have implications for the hunt for 'fat genes'. Although some genes may be linked to the risk of getting fat, obesity is less the result of individual genetic propensities than of the shift in environmental conditions, says Ricardo Uauy, an expert on public health and nutrition at the London School of Hygiene and Tropical Medicine. "The past 50 years is too short to modify our evolutionary trajectory."

Declan Butler

Do we have enough of the right kind of food for everyone on the planet?

Certainly if one considers the world's chronically undernourished, who now number some 850 million people⁵, the answer is no. But the surprise is that many of those who are better off or who live in countries with abundant food supplies still fail to get the nutrients they need, and may even be overweight in spite of this.

The problem of chronic hunger occurs almost entirely in poor countries. The condi-

tion is particularly lethal to children, of whom more than 3.7 million died in 2002 from the health consequences of being underweight. Another estimated 850,000 died because their diet—although sufficiently rich in calories—did not contain enough vital components such as iron, vitamin A and zinc⁵.

That lack of proper nutrients is also a phenomenon in wealthy countries, where food insecurity, if not starvation, is surprisingly common. In the United States, for example, 12.6 million households (about 11%) fall short of basic food needs at some point during the year. In about a quarter of those cases, people fail to get government food aid or find private charities to make up the difference, and so go hungry⁶.

Ironically, poverty and obesity often go hand in hand in developed countries⁷. "Obesity is a disease of the poor in rich countries, whereas in poor countries obesity is a disease of the rich," says Katan.

A key factor is that junk food tends to offer the most calories for the least money. "This is the single most important factor influencing food choice," says Marion Nestle, professor of nutrition,

food studies and public health at New York University. In US supermarkets, for instance, a 270-calorie doughnut costs about 75¢, the same as 125-calorie apple.

The ready availability of processed and fast foods in many corners of the globe is now making them the natural choice, particularly for the poor and uneducated. Apart from being cheap, they have a natural appeal, Nestle says. "Eating highly refined food rich in sugar and fat is a kind of joy, which poor people do not frequently have," she says.

Although it is true that wholesome foods are also available throughout the industrialized world, evidence suggests that even slight inconvenience is enough to put people off buying them. A representative study involving participants in the US food-stamp programme shows that people tend to buy more fruit the closer they live to a supermarket⁸.

Achim Schneider



1. Trichopoulos, D. & Lagiou, P. *Public Health Nutr.* 7, 949–951 (2004).

2. Padez, C., Fernandes, T., Mourão, I., Moreira, P. & Rosado, V. *Am. J. Hum. Biol.* 16, 670–678 (2004).

3. Puska, P. *Acta Cardiol.* 55, 213–220 (2000).

4. Leonard, W. R. et al. *Annu. Rev. Anthropol.* 34, (in the press).

5. *The State of Food Insecurity in the World 2004* (UN Food and Agriculture Organization, Rome, Italy, 2004); available at: <http://ftp.fao.org/docrep/fao/007/y5650e/y5650e00.pdf>.

6. Nord, M., Andrews, M. & Carlson, S. *Food Assistance and Nutrition Research Report no. 42: Household Food Security in the United States, 2003* (USDA Economic Research Service, Washington DC, 2004); available at: www.ers.usda.gov/publications/fanrr42.

7. Drewnowski, A. & Specter, S. E. *Am. J. Clin. Nutr.* 79, 6–16 (2004).

8. Rose, D. & Richards, R. *Public Health Nutr.* 7, 1081–1088 (2004).

Vital resource should be open to all physicists

Putting control in the hands of a few can enforce orthodoxy and stifle innovative ideas.

Sir — Your News story “Rejected physicists instigate anti-arXiv site” (*Nature* **432**, 428–429; 2004) reports a response from Paul Ginsparg, the founder of the preprint server arXiv.org, to criticisms of its publication policies. Ginsparg states that the rules governing who can and cannot publish are clearly stated on the site, and that the archive is designed for “communication among research professionals, not as a mechanism for outsiders to communicate to that community”.

The cases documented by myself and others on the ArchiveFreedom website show that there is more to the story.

The exclusion of particular individuals and particular ideas from arXiv appears to me to be deliberate. If a rule can be invoked in support, however tenuous the link, the rule is quoted; otherwise, submissions are simply ‘deleted as inappropriate’. For example, having stated that a very distinguished physicist’s strong support of a submission carried no weight because this physicist “was not intimately familiar with the work in question”, the moderators simply ignored subsequent support from an endorser with publications on the same subject.

In another example, the moderators’ response to the information that more

than one eminent physicist had an interest in a subject that they wished to bar was: “We are always thrilled to hear when people find an avocation that keeps them off the streets and out of trouble.”

ArXiv has become a vital communicative resource for the physics community. The moderators’ attitude to any challenge to conventional thinking is likely to result in the loss to science of important innovative ideas. Radical changes are required in the way the archive is administered.

Brian D. Josephson

Department of Physics,
University of Cambridge,
Cambridge CB3 0HE, UK

Climate blog could score with newer hockey stick

Sir — In your Editorial “Welcome climate bloggers” and News story “Climatologists get real over global warming” (*Nature* **432**, 933 and 937; 2004), the newly created RealClimate blog (www.realclimate.org) is introduced as a website battling distorted media coverage on global-warming research.

As a member of the climate-research community and inspired by your enthusiastic introduction, I navigated RealClimate with high expectations. I was, however, sadly disappointed by a posting by M. E. Mann on 4 December: “Temperature variations in past centuries and the so-called ‘Hockey Stick.’” Among other data, this included an overview of temperature change during the past millennium as reconstructed by various climate proxies including borehole data.

I cannot comment on the accuracy of the rest of the posting, but I was concerned to find that Figure 1, showing temperature change in the Northern Hemisphere, included an outdated and erroneous reconstruction of borehole data by M. E. Mann *et al.* (*J. Geophys. Res.* **108**, 4203; 2003). In my view, the website should have used a later version (S. Rutherford and M. E. Mann, *J. Geophys. Res.* **109**, D11107; 2004), which acknowledged an error in the earlier paper, or other more recent reconstructions. To be fair, the authors of the website added a correction after I drew their attention to this.

Your Editorial asserts that there is little reason to doubt that RealClimate’s goal, “to provide solid scientific comment to journalists and other interested parties”, can be reached. But you also warn of the dangers of “a rapid-rebuttal service,

run with minimal peer review”. Let us hope that future postings on RealClimate will fulfil our high expectations.

Shaopeng Huang

Department of Geological Sciences, University of
Michigan, Ann Arbor, Michigan 48109-1063, USA

Best way to protect rock art is to leave it alone

Sir — Your News Feature about the Lascaux cave, “The film crew” (*Nature* **433**, 100–101; 2005), expresses the hope that, thanks to new technologies, “tourists will get their chance to see the real version of this ancient site”. Unfortunately, the history of Lascaux’s conservation since its discovery in 1940 leads one to expect the opposite. Today, the cave is closed even to experts in rock art.

Contrary to your report, the cave’s climate did not return to its original state after the 1963 closure. How could it? Several hundred cubic metres of soil had been removed between 1940 and 1958, when an air-conditioning system was installed. The green algae, mosses and bacteria were eventually eradicated from the walls, but the growth of small white calcite crystals on the surface caused by rises in temperature, humidity and carbonic gas associated with visitors to the caves could only be stopped, not reversed.

Lascaux remains extremely fragile. The 2001 invasion of the floor by white *Fusarium* fungi and the (fortunately limited) appearance of black spots on the ceiling remind us of the dangers of interfering with rock-art sites. The best technology for preserving them is to leave them alone.

Luc Allemand*, **Paul G. Bahn†**

*La Recherche, 4 rue du Texel, 75014 Paris, France
†428 Anlaby Road, Hull HU3 6QP, UK

Online submission makes authors do all the work

Sir — During the past two years, most of the leading scientific journals have switched to an electronic system for manuscript submission.

Superficially, this might seem to be progress. But in practice, the onus for the preparation of publication-quality manuscripts, particularly figures, has quietly switched from the journals to the authors, who are now expected to run mini-desktop publishing operations from their offices and laboratories.

The submission of a paper can now take days of fiddling with various computer programs, where once it took a few minutes to print a manuscript and shove it in an envelope. The end-product is no better, nor is publication any quicker, and page charges (for those journals operating that system) are as high as ever.

So just who has benefited, or profited, from the change? The authors or the journals? And how do scientists in less-developed countries with older computer systems cope with the quixotic demands of the electronic systems?

The degree of user-unfriendliness varies from journal to journal (*Nature*’s is far from the worst), but avoiding the most hassle-associated systems is now, in my case at least, a significant factor to be taken into account when choosing a journal for submission of a paper.

John P. Moore

Department of Microbiology and Immunology,
Joan and Sanford I. Weill Medical College of
Cornell University,
1300 York Avenue, W-805 New York,
New York 10021, USA

Scandals and safeguards

Is scientific fraud on the increase?

The Great Betrayal: Fraud in Science

by Horace Freeland Judson

Harcourt: 2004. 463 pp. \$28

Daniel S. Greenberg

The scientific enterprise is unquestionably afflicted by ethical, financial and bureaucratic woes, as often reported in *Nature* and elsewhere. But these problems are far worse than most of us realize, according to Horace Freeland Judson in *The Great Betrayal*, a brazen indictment of the condition of contemporary science.

Among scientists, the theft of intellectual property is “epidemic”, Judson contends, and the enshrined processes of peer review for grants and publication have been rendered “moribund” by politics, cronyism and deceit. Furthermore, he asserts, the transition in research from healthy financial growth to a steady state is intensifying the difficulties. Judson acknowledges that the evidence for these stark assertions is scanty, because, like all clandestine, deviant behaviour, it is hard to measure precisely. “We have not yet found a way of getting at the true incidence of fraud in science,” he observes.

No matter. Taking a tip-of-the-iceberg approach, Judson extrapolates from scores of documented episodes in the pantheon of scientific fakery, many of them also recounted in a 1983 book of similar title, scope and dour conclusions, *Betrayers of the Truth* by William Broad and Nicholas Wade (Simon & Schuster). Judson revisits the hoary Pildown hoax, the fakery mill that flourished in a prestigious cardiology laboratory at Harvard Medical School 25 years ago, and the fraudulent tissue-transplant reports that roiled the Sloan-Kettering Institute in the 1970s, along with others of comparable infamy.

Bringing the roll of dishonour up to date, Judson concludes that the rot is not merely episodic and occasional, but runs wide and deep. It is not, he says, the rarity supposed by Daniel Koshland when, as editor of *Science* in 1987, he brashly wrote that “99.9999% of [scientific] reports are accurate and truthful”. Psychopathology — the establishment aetiology for scientific misdeeds — is not the primary factor, Judson argues. Rather, the disorder is integral to modern science, inexorably arising from inadequate mentoring, veneration of high-volume publication, chases for grants and glory, political pressures for practical results, and insufficient budgets that inspire ethical shortcuts.

Along the way, Judson fires salvos of derision at David Baltimore, best known to the public not for his Nobel prize but for



David Baltimore was embroiled in controversy when he defended a colleague accused of misconduct.

his tenacious, controversial defence of a research collaborator who was accused of misconduct but officially exonerated after a decade of government inquiries. It was the Baltimore case, Judson explains, that drew him to the trail of fraud in 1991. In this, the book that ensued, Judson gives Baltimore the lengthiest, most detailed attention, and even tells us that the Rockefeller University faculty “found the data in his proffered dissertation of borderline quality at best, thin.”

Judson draws heavily on the literature of scientific delinquency. But curiously he makes no reference to the definitive work, *The Baltimore Case* (W. W. Norton, 1998) by Daniel J. Kevles, although Judson’s book contrasts sharply with Kevles’ exoneration of Baltimore. Judson, as others have before, charges Baltimore with arrogance and making misleading allegations of political interference in science. Baltimore, he states, could have ended the controversy at an early stage “by scrutinizing the disputed data and announcing that he was reconsidering the paper. This he refused to do.”

From the Baltimore case and other eruptions, old and new, Judson infers that dangerous pathologies infest the culture of science. He says they have been little touched by government-mandated safeguards in recent years that call for the ethical tutoring of graduate students, protection of whistle-blowers, retention of laboratory records, and systematic enquiry into fraud allegations. More important than the guilt or innocence of

individuals, Judson insists, “is the protection of the scientific process and of the integrity of the scientific record”. These, he says, are increasingly neglected values in the intensely competitive world of modern science.

Judson certainly merits attention. A scholar and journalist with a wide acquaintanceship in the international scientific community, he is the author of a highly respected book, *The Eighth Day of Creation* (Simon & Schuster, 1979), and is the founder and former director of the Center for History of Recent Science at The George Washington University.

His arguments, however, strike me as being far-fetched, dated and poorly aimed. Fraud in science, to the extent that it is calculable, seems to be no worse today than in previous times. It has perhaps been checked to some extent by the aforementioned safeguards and, as Judson notes, by the power of the Internet to detect plagiarism of text, if not of ideas. Steady-state funding may pose dangers, but current annual US government spending on biomedical research has risen to nearly \$30 billion, up from \$12 billion in 1996, when Judson and others bemoaned what they saw as an impending steady state. Meanwhile, California and other states are planning to spend large sums on stem-cell research and other areas of biotechnology.

The main threat to scientific purity today originates in corporate money and wiles aimed at co-opting the good name of science for the pursuit of profit, as revealed in recent

B. W. SMITH/TIME LIFE PICTURES/GETTY IMAGES

pharmaceutical scandals. Withholding of clinical research data unfavourable to pharmaceutical products, concealment of financial interests in drug trials, ghosted papers for the promotion of drugs, and lucrative consulting deals for academic and medical 'thought leaders' are among the techniques that have surfaced. As former Harvard president Derek Bok laments in *Universities in the Marketplace* (Princeton University Press, 2003): "Most universities have not done all they should to protect the integrity of research. Many have not even shown they are seriously concerned about doing so."

Of these threats to the well-being of science, Judson says virtually nothing. ■

Daniel S. Greenberg is a guest scholar at the Brookings Institution, Washington DC, USA. He is the author of *Science, Money, and Politics*.



N. PISARENKO/AP

The sweet taste of success: honey has a wide range of culinary and medicinal uses.

A scientific feast

On Food and Cooking: The Science and Lore of the Kitchen, 2nd edition

by Harold McGee

Scribner/Hodder & Stoughton: 2004. 896 pp. \$35, £30

Hervé This

When *On Food And Cooking* was first published in 1984 it became a best-seller in English-speaking countries, and deservedly so. This was after Nicholas Kurti and I began our first experiments in the kitchen but before we coined the phrase 'molecular gastronomy' — the discipline now has its own national and international workshops, conferences, courses and seminars. Molecular gastronomy isn't just concerned with cooking; it is the part of food science that relates to gastronomy in general. According to the French gastronome Jean-Anthelme Brillat-Savarin: "Gastronomy encompasses all knowledge about man as he is eating."

Food science has a long history. Some 2,000 years ago, the anonymous author of the London papyrus used a balance to find out whether fermented meat weighed less than fresh meat, because of some 'emanation' that is lost. Much later, Antoine Parmentier, Michel-Eugène Chevreul, Benjamin Thomson (Count Rumford), Emil Fischer and others made remarkable contributions. Their work became very popular. For example, around the time of the Second World War, Edouard de Pomiane, a biologist at the Pasteur Institute, was writing best-sellers on food science in France.

In the same tradition, Harold McGee started publishing his books. I confess that I am not an impartial judge: he is a friend and one of the core participants of a series of international workshops on molecular gastronomy that I have organized every two

years or so at the Ettore Majorana Centre for Scientific Culture in Erice, Sicily, since 1992.

What kind of information will a reader will find in his book? In the arbitrarily chosen chapter 12, about sugars, chocolate and confectionery, there are sections on the history and nature of sugars; sugars and syrups; confectionery; and chocolate. In the section on sugars and syrups, we learn about honeybees and how they make honey (gathering nectar and transforming it) before finding out about processing and storing honey, its flavour, using it in cooking, honey and health, and infant botulism. There are boxes on the advance of the bee in North America and sweet ants, along with a picture of a honeycomb and a sketch of honey. As you can imagine, the book is a natural history of food and cooking.

The publisher's claim that this new edition is completely revised and updated really is true. I now understand why McGee has been so busy for the past few years: the new edition has nearly 900 pages full of detailed information on food, its production and its transformation during cooking. Not only has the book's physical appearance changed, but the table of contents has been extended and the text greatly modified. For instance, the first edition explained that the avocado "has been cultivated in Central America for perhaps 7,000 years", but the same section now begins by explaining that "the avocado tree *Persea americana* is a native of Central America and a member of the laurel family, a relative of the bay laurel, California bay, and sassafras". Instead of reading that "its fat content, at 20%, is about 20 times the average for other fruits", we now learn that "avocado fruits are remarkable for containing little or no sugar or starch, and for being as much as 30% oil, the equivalent of well-marbled meat".

The book seems to be oriented mainly at cooks, probably because so many used the

first edition. The text has been divided into small, easy-to-swallow sections; there are probably more cooking tips (but no recipes) than in the first edition; and the pictures are much improved (although they are in black and white). The scientific descriptions and explanations have been sharpened, and the references have been removed from the text and figure legends, and grouped together.

Of course, it is possible to criticize the book, but the reason for any imprecision is probably the lack of space: McGee had to distil the information and has given only the most useful. For example, he explains that astringency "is caused by a group of phenolic compounds consisting of 3 to 5 carbon rings, which are just the right size to span two or more normally separate protein molecules, bond them and hold them together". However, it is generally accepted that tannins with a relative molecular mass of about 5,000 could contribute to astringency. In addition, monomeric flavanols and some other simple phenolics such as gallic acid, which are not chemically defined as tannins, also precipitate proteins and are perceived as astringent. Finally, procyanidins become gradually less bitter and more astringent as their relative molecular mass increases. But then, I am fascinated by chemistry, whereas McGee was trying to summarize a wealth of information to provide only the main point for cooks.

Anyone seeking the scientific details could complement this book with the latest edition of the remarkable *Food Chemistry* by H. D. Belitz *et al.* (Springer, 2004). That new editions of these books have been published at the same time is probably no coincidence — molecular gastronomy is fashionable now in culinary as well as scientific circles. ■

Hervé This is in the INRA Group on Molecular Gastronomy, Laboratory of Chemistry, Collège de France, 11 Place Marcelin Berthelot, 75005 Paris, France.

Looking ahead to future brain studies

The New Brain Sciences: Perils and Prospects

edited by Dai Rees & Steven Rose
Cambridge University Press: 2004. 316 pp.
£65, \$120 (hbk); £24.99, \$43 (pbk)

David Papineau

The brain sciences are undoubtedly burgeoning. In recent years, molecular biochemistry and genomic mapping have been combining with fancy functional imaging techniques to tell us more and more about the brain's internal workings.

Still, is all this new knowledge really such a good thing? The contributors to *The New Brain Sciences* are not sure. As Steven Rose explains in the introduction: "You will find no gung-ho overoptimistic forecasts of the wondrous cornucopia of benefits that neuroscience might bring here." True, he immediately adds: "Nor, though, are our authors doom-sayers with an almost automatic rejectionism in response to new findings." However, even if they aren't automatic rejectionists, most of the authors certainly seem to be worried about something.

In the end, though, this book is rather reassuring. The general tenor of the essays is that there is nothing in the new brain sciences to overturn anything we hold dear. Only muddled thinking, the contributors say, could make you suppose that neuroscience is going to radically alter our lives.

The articles are derived from a pair of recent conferences, and are divided into three sections. The first part asks whether we are more than the sum of our biochemical parts; the second considers whether biochemical determinism means that we are not authors of our own actions; and the final section wonders whether neuroscientific advances will lead to new medical techniques.

The contributors to the first section are emphatically of the anti-reductionist party. The philosopher Mary Midgley parades once more in her familiar colours, stressing that there is more to human nature than can be gleaned from the workings of neurotransmitters. Her message is echoed by the evolutionary psychologist Merlin Donald and the sociologist Hilary Rose. These authors are of course quite right, but I did wonder who they took their opposition to be. Poor Richard Dawkins comes in for some flak, but it's not exactly clear what for. It might also have been helpful if these essays had distinguished more clearly between the uncontroversial methodological point that theories outside neuroscience can help us to understand people and the rather more controversial metaphysical issue of whether we are made of anything more than molecules.

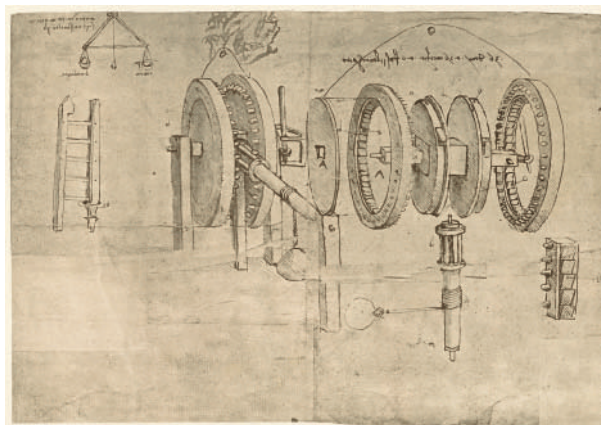
Exhibition

Leonardo's legacy

The *Codex Atlanticus* is a collection of more than a thousand sheets of the scientific and technical drawings of Leonardo da Vinci, put together at the end of the sixteenth century by Pompeo Leoni, a sculptor. Leoni was trying to organize Leonardo's work into categories, cutting and pasting drawings from original notebooks on to the atlas-sized pages that give the *Codex Atlanticus* its name. Some drawings were damaged in the process and others lost.

History continued to be unkind to the drawings. The codex was appropriated by Napoleon at the end of the eighteenth century, before being returned from Paris to Milan in the mid-nineteenth century. Early photographers, most fortunately as it turned out, captured images of the sheets on huge glass plates, and these formed the basis of a luxurious reproduction in 1906 of the entire codex. The original sheets were poorly restored in the 1950s and the early photographs are more valuable to historians than the original sheets.

Copies of the 1906 edition — a collaboration



between the Accademia dei Lincei, Italy's national academy, and the publisher Anthelios — are now rare. But one forms the centrepiece of an exhibition currently on view in the austere, baroque rooms of the Accademia dei Lincei in Rome. The sheets are displayed alongside modern interpretations, or counterparts, of the drawings: there is a reconstruction of Leonardo's helicopter in wood and a Ferrari engine, for example.

In mid-March the exhibition begins an extensive European tour, taking in Budapest, Bratislava, Warsaw, Bolzano and other European cities in 2006, before moving to the United States and Japan in 2007.

A.A.

The second section is particularly concerned with responsibility in a legal context. If it turns out that someone was predisposed to commit a crime because of some genetically determined feature of their brain, should they be punished? All the contributors agree that this wouldn't necessarily be a good excuse. For better or worse, we currently hold people responsible for their choices, provided that they are capable of deliberation. If they have the bad luck to be lumbered with a nefarious nature, we expect them to curb it: paedophiles aren't held to account merely for being attracted to children, but rather for succumbing to their desires. This doesn't change just because your criminal tendencies are foisted on you by your genes. If you are capable of deliberation, it's still up to you whether or not you give in to those tendencies.

This theme is repeated in a series of lucid articles from the philosopher Peter Lipton, the ethologist Patrick Bateson, the appeals-court judge Stephen Sedley and the academic medical lawyer Alexander McCall Smith. Those who know of the latter only from his fictional chronicles of the No. 1 Ladies' Detective Agency in Botswana will not be surprised by the elegant prose with which he carries out his day job.

By and large, the third section is not particularly enthusiastic about the medical

promise of recent neuroscientific advances. But even here the pessimism is diluted. Alongside articles reminding us about the poor track record of neurosurgical intervention, and about the incoherence of much of the work on genes and intelligence, there is a markedly calm discussion by Paul Cooper of the role that the drug Ritalin (methylphenidate) can play in treating attention-deficit/hyperactivity disorder. There are also two useful articles analysing the ethics of stem-cell research and the prospects for resulting therapies. An insightful contribution by David Healy explains the techniques used by pharmaceutical companies to market their wares, and left me hungry for more on the way that commercial imperatives are distorting the development of new drugs.

Overall, this volume does much to combat various kinds of bad reductionist thinking. But it does little to show that the 'new brain sciences' pose any particular threat to anything. Prospective readers should also be warned that there is scarcely any information about the brain sciences themselves. Still, there is no harm in being reminded once more that there is more to life than basic scientific knowledge.

David Papineau is professor of the philosophy of science, King's College London, The Strand, London WC2R 2LS, UK. His latest book is *Thinking about Consciousness*.

Schrödinger's mousetrap

Part 6: A cryptic response.

**Ilana Goldhaber-Gordon
and David Goldhaber-Gordon**

"Veronique Dubois?"

"And you must be the inspector investigating Rufus Jaeger's death, no? I was wondering if you would want to talk to me." Veronique Dubois and Karl Lister settled themselves into a small lecture room and Lister began his questioning, gently at first.

"I'd like to hear your impressions of Professor Jaeger," he said.

Dubois smiled. "I liked Rufus. You must know by now that not everyone felt that way. Rufus could be quite a ... a shark. But," a contemplative pause, "Rufus and I understood each other." She shook her head and sighed. "I can't believe he's gone, so quickly. You don't really believe it was murder, do you?"

"We have to consider that possibility. Tell me, how often did you see Professor Jaeger?"

"Not very often. We'd attend a few of the same conferences each year, and each time we'd share some meals. But we wouldn't seek each other out. There isn't much to talk about if you can't discuss science."

"You can't discuss science?"

"No. Rufus and I are, were, competitors."

"I see," Lister said, although he didn't really. "Do you talk science with Professor Pruszczyński?" Phew, he'd pronounced it perfectly.

"Of course," Dubois said condescendingly. "Petra and I are collaborators."

"I see," Lister said again. "And how do you determine who should be a collaborator and who a competitor?"

Dubois looked confused. "I didn't always think of Rufus as a competitor," she said, embarrassed by her bitter tone even as the words left her mouth.

"Did something happen to change your perspective?"

"Early on, I was a bit foolish. I let my excitement betray me, and on a visit to Rufus's lab I shared many ideas. Then an idea I had been nurturing appeared in one of their papers, and I was never acknowledged." During this rush of words, Lister suddenly became conscious of Dubois' accent.

"That must have been frustrating," he said.

"Yes. No. I mean, of course it was frustrating at the time. But it was a relatively cheap lesson. I needed to learn to ... er ... play hardball, and now I know. But, getting back to the present, I think you should see this." Dubois quickly pulled out her palm pilot, stroked in a few commands and thrust it at Lister. "It's a note Wilfred sent to Ludmilla."



Lister read:

hi. I think you should be careful about your relationship with RJ. He is a liar and a cheat. an example. He has just accepted a prize of for the paper on parallel quantum codes which is entirely my work. I realize now that I should have asked the journal from the list of authors

"And how did this message come to be in your hands?" Lister asked.

"As part of my research on quantum cryptography," Dubois responded smoothly. "Both Rufus and I are developing quantum e-mail systems and this message was sent within Rufus's network. To test my techniques, I've been trying to intercept quantum encrypted e-mail messages. You may know that quantum cryptography is supposedly unbreakable. An eavesdropper reveals herself as soon as she attempts to intercept the key..."

"The key?"

"The secret code used to convert a seemingly nonsensical message into the true message. A quantum key is developed collaboratively, between sender and recipient, and it cannot be intercepted without the third party revealing herself."

"Then how..."

"How did I intercept this message? The security of these communications relies on the sender and recipient properly identifying one another. If an eavesdropper could impersonate the recipient, she might fool the sender into developing a quantum key with her, the eavesdropper, rather than with the proper recipient. Which is what I did. So, what do you think of the note?"

"The wording seems a bit strange. Are there words missing?"

Dubois looked at him sharply. "There are, in fact. This is a strange thing: Rufus's net-

work breaks up each 'q-mail' message into 20 or more shorter messages. They construct a new, independent key for every one to two words. From my perspective, this means we rarely intercept the key—or keys, I should say—for an entire message, but only for pieces of it. We can only read messages in fragments. It's unconventional what they are doing, and I had assumed it was because they are new to the quantum-cryptography game. But you know, I wonder. We can read only fragments, and the proper recipient can read only the fragments we miss. Perhaps this breaking up of messages is an extra safeguard to alert against intruders..."

"So Dr Shlomiuka read only the words that are missing from this text?"

"Yes, and whether or not she was aware that she was missing words depends on how her encryption software interfaces with the user."

"I see," Lister scanned the note again.

"It's sad, isn't it?" Dubois clicked her tongue. "Wilfred's been passed over again and again for promotion. And he's been admiring Ludmilla for years and getting nowhere with her. And he clearly has—or had—a grudge against his boss, Rufus. You know, if I were you, I would have a talk with him."

"Yes," Lister said, fingering an imaginary cigarette. "Thanks for the tip."

"My pleasure. Well, if that's all?" Dubois was looking at him expectantly now, suddenly ready to end the discussion. Lister had a nagging feeling that he was missing something, but Dubois' expression discouraged prolonging the conversation.

"Just one more thing, Mi..., er, Professor Dubois," Lister said. "What were you doing during the coffee break before the session?"

Dubois smiled at the classic question. "Petra and I were in a study room, discussing our collaboration, as I'm sure she told you." She paused, then looked the inspector in the eye. "Mr Lister, Petra Pruszczyński and Rufus Jaeger were scientific competitors, not jealous lovers. Theirs was not the sort of competition that provokes murder."

To be continued...

Ilana Goldhaber-Gordon is the author of forthcoming textbook Biochemistry for the Biology Graduate Student. David Goldhaber-Gordon is co-director of the NSF-Stanford-IBM Center for Probing the Nanoscale and is in the Department of Physics, Stanford University, Stanford, California 94305, USA.

Who do you think killed Rufus Jaeger? Catch up on all the evidence and vote for your suspect at
www.nature.com/news/mousetrap

Elephant breakdown

Social trauma: early disruption of attachment can affect the physiology, behaviour and culture of animals and humans over generations.

G. A. Bradshaw, Allan N. Schore, Janine L. Brown, Joyce H. Poole and Cynthia J. Moss

The air explodes with the sound of high-powered rifles and the startled infant watches his family fall to the ground, the image seared into his memory. He and other orphans are then transported to distant locales to start new lives. Ten years later, the teenaged orphans begin a killing rampage, leaving more than a hundred victims.

A scene describing post-traumatic stress disorder (PTSD) in Kosovo or Rwanda? The similarities are striking — but here, the teenagers are young elephants and the victims, rhinoceroses. In the past, animal studies have been used to make inferences about human behaviour. Now, studies of human PTSD can be instructive in understanding how violence also affects elephant culture.

Psychobiological trauma in humans is increasingly encountered as a legacy of war and socio-ecological disruptions. Trauma affects society directly through an individual's experience, and indirectly through social transmission and the collapse of traditional social structures. Long-term studies show that although many individuals survive, they may face a lifelong struggle with depression, suicide or behavioural dysfunctions. In addition, their children and families can exhibit similar symptoms, including domestic violence. Trauma can define a culture.

How PTSD manifests has long been a puzzle, but researchers today have a better idea as to why the effects of violence persist so long after the event. Studies on animals and human genocide survivors indicate that trauma early in life has lasting psychophysiological effects on brain and behaviour.

Under normal conditions, early mother–infant interactions facilitate the development of self-regulatory structures located in the cortic limbic region of the brain's right hemisphere. But with trauma, an enduring right-brain dysfunction can develop, creating a vulnerability to PTSD and a predisposition to violence in adulthood. Profound disruptions to the attachment bonding process, such as maternal separation, deprivation or trauma, can upset psychobiological and neurochemical regulation in the developing brain, leading to abnormal neurogenesis, synaptogenesis and neurochemical differentiation. The absence of compensatory social structures, such as older generations, can also impede recovery.

Elephant society in Africa has been



Social bonds guide an elephant's development.

decimated by mass deaths and social breakdown from poaching, culls and habitat loss. From an estimated ten million elephants in the early 1900s, there are only half a million left today. Wild elephants are displaying symptoms associated with human PTSD: abnormal startle response, depression, unpredictable asocial behaviour and hyperaggression.

Elephants are renowned for their close relationships. Young elephants are reared in a matriarchal society, embedded in complex layers of extended family. Culls and illegal poaching have fragmented these patterns of social attachment by eliminating the supportive stratum of the matriarch and older female caretakers (allomothers).

Calves witnessing culls and those raised by young, inexperienced mothers are high-risk candidates for later disorders, including an inability to regulate stress-reactive aggressive states. Even the fetuses of young pregnant females can be affected by pre-natal stress during culls. The rhinoceros-killing males may have been particularly vulnerable to the effects of pre- and postnatal stress for two reasons. Studies on a variety of species indicate that male mammalian brains develop at a slower rate relative to females, but also that elephant males require a second distinct phase of socialization.

As with females, male socialization begins during infancy with the mother and a tight constellation of allomothers. But in adolescence, males leave the natal family to participate in older all-male groups, a period coincident with a second major stage of brain reorganization identified in humans. Cull orphans sustain a series of traumas, such as premature weaning, shock and the lack of older male socialization. The critical role of older males in normal social development

was clearly demonstrated when researchers re-introduced older bulls to quell the young males' violence. Hyperaggression and abnormally early musth cycles (periods of sexual activity and hormonal shifts) both ceased.

Elephant hyperaggression is not an isolated event. At another heavily affected African park, intraspecific mortality among male elephants accounts for nearly 90% of all male deaths, compared with 6% in relatively unstressed communities. Elsewhere, including Asia, there are reports of poor mothering skills, infant rejection, increased 'problem animals' and elevated stress-hormone levels.

Elephant sociality is both a strength and a weakness. As with humans, an intact, functioning social order helps buffer trauma. But as human populations increase, more elephants are likely to live in environments characterized by severe anthropogenic disturbance. Current methods for conserving both wild and captive elephant populations fail to preserve elephant social systems. Even successful rehabilitation centres, such as The David Sheldrick Wildlife Trust, can only partially restore social processes because there are not enough older herd members. There is an added danger to social breakdown, namely that selection for asocial heritable traits in the absence of normal socialization may increase under adverse conditions. All these factors bring into question what kinds of behaviour are being promulgated in both *ex situ* and *in situ* conservation programmes, and compel new conservation strategies that promote normal social patterns.

Neuroscience has demonstrated that all mammals share a ubiquitous developmental attachment mechanism and a common stress-regulating neurophysiology. Now, a wealth of human–animal studies and the experiences of human victims of violence are available to help elephants and other species survive. ■

G. A. Bradshaw is at the Environmental Sciences Graduate Programme and the Department of Forest Science, Oregon State University, Oregon, USA.

Allan N. Schore is in the Department of Psychiatry, University of California, Los Angeles, California, USA.

Janine L. Brown is at the Smithsonian National Zoological Park, Washington DC, USA. Joyce H. Poole and Cynthia J. Moss are at the Amboseli Elephant Research Project (AERP), Nairobi, Kenya.

FURTHER READING

Clubb, R. & Mason, G. *A Review of the Welfare of Elephants in European Zoos* (RSPCA, Horsham, 2002).
Schore, A. N. *Affect Dysregulation and Disorders of the Self* (W. W. Norton, New York, 2003).
Slotow, R. et al. *Nature* **408**, 425–426 (2000).

Snow maker for the ice ages

Katharina Billups

In the Northern Hemisphere, large-scale glaciation was initiated comparatively recently. Paradoxically, it seems that the trigger was a seasonal warming of the sea surface in an upwind oceanic region.

Over the past 50 million years, the Earth's climate has been cooling (Fig. 1). Although Antarctica has been glaciated for at least the past 35 million years¹, large ice sheets did not appear in the Northern Hemisphere until about 2.7 million years ago. Earth scientists largely agree that overall climate cooling is associated with decreasing levels of carbon dioxide in the atmosphere^{2,3}, and that ice sheets can only grow if sufficient moisture is available and winter snow survives the summer heat⁴. But what triggered the onset of the ice ages 2.7 million years ago? Explanations have focused on continental temperatures^{4,5}, with identification of potential moisture sources from the Atlantic^{5,6}, but there remain many open questions⁷.

Haug *et al.*⁸ (page 821 of this issue) contribute an important piece to the ice-age puzzle. Geochemical evidence suggests that, 2.7 million years ago, the seasonal temperature contrast of the subarctic Pacific Ocean sea surface became larger as summers warmed and winters cooled. Warmer summer sea-surface temperatures result in a warmer atmosphere that can hold more moisture. Like a snow gun blasting away at ski slopes, westerly winds blow the moisture onto the cold North American continent where it falls as snow and accumulates as ice (Fig. 1, inset; Fig. 2, overleaf).

Haug *et al.* have combined geochemical expertise with numerical modelling to present an integrated approach to the origin of the ice ages. Evidence comes from the floor of the subarctic Pacific Ocean, on which the remains of certain species of marine plankton (diatoms, coccolithophores and foraminifera) have accumulated over time. The primary evidence for summertime warming 2.7 million years ago stems from the biochemistry of coccolithophores, which varies according to temperature⁹. Augmenting this well-established index are the $^{18}\text{O}/^{16}\text{O}$ ratios in the siliceous tests of diatoms, a comparatively more complex measure of palaeotemperatures¹⁰.

At first glance, the results from these two recorders contrast with other climate indicators in this region. Foraminiferal $^{18}\text{O}/^{16}\text{O}$ ratios — a classical indicator¹¹ — from the same deep-sea sediments suggest sea-surface cooling 2.7 million years ago. This particular evidence is corroborated by perhaps the most intuitive indicator of climatic cooling,

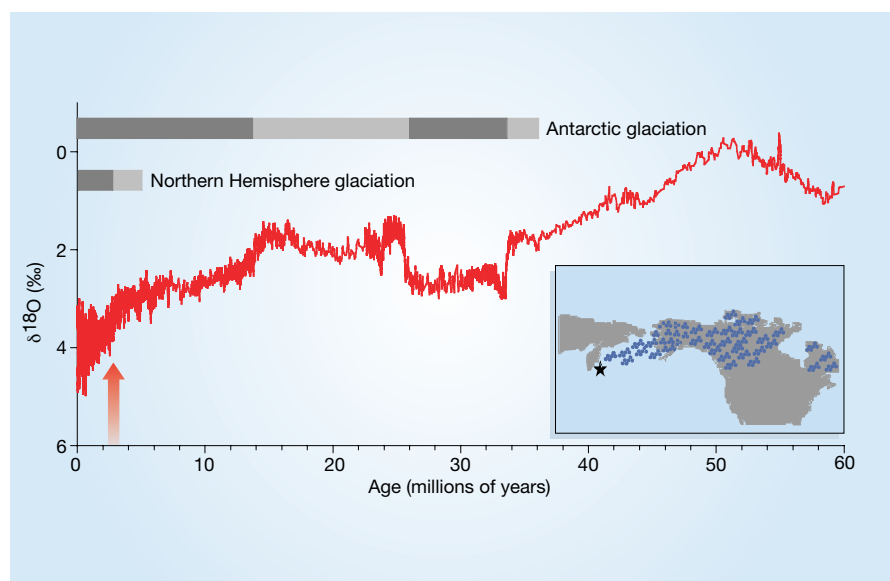


Figure 1 Global climate change over the past 60 million years. This record, showing a mainly cooling trend, is inferred from foraminiferal oxygen-isotope records from all major ocean basins¹ with $^{18}\text{O}/^{16}\text{O}$ ratios plotted as per mil deviation from a standard ($\delta^{18}\text{O}$). The horizontal grey bars indicate the relative extent of polar ice sheets — light grey, ice volumes less than half of the maximum extent; dark grey, ice volumes close to the maximum extent (after ref. 1). Haug *et al.*⁸ provide evidence that the initiation of the Northern Hemisphere ice ages, 2–3 million years ago (arrow), was linked to the development of a stratified sea surface in the subarctic Pacific, which resulted in warmer sea-surface temperatures in late summer. Inset, the warmer sea surface was a source of moisture for the overlying atmosphere, with westerly winds loading the snow gun that produced large ice sheets on the North American continent. The star indicates the authors' study site⁸.

an increase in the amount of debris of continental origin delivered to the site by icebergs.

How can these apparently contradictory observations be reconciled? Haug *et al.*⁸ point to seasonal changes in the biological communities of the subarctic Pacific Ocean where, in modern times, different plankton communities populate the various seasons. Coccolithophores and those species of diatoms used for the geochemical analyses prefer the warm ocean surface of late summer and autumn. The particular foraminiferal species used for analysis, on the other hand, are more prolific during late winter and spring when the sea surface is fertile due to mixing with deeper, colder, nutrient-rich water.

Thus the two seemingly opposing temperature trends 2.7 million years ago simply reflect an increase in seasonality in the subarctic Pacific Ocean, which is consistent with other reconstructions of events in the North Pacific Ocean⁷. This then provides the configuration on which to build an ice age:

late winter cooling reflects climate cooling, allowing snow to accumulate; late summer warming increases the atmosphere's potential to hold moisture and to load the snow gun (Fig. 1, inset).

What, then, caused the sudden increase in late summer temperatures? To answer this question, Haug *et al.*⁸ refer to the physical properties of sea water itself. Water has a high heat capacity, which means that the surface ocean remains warm long after overlying air and adjacent land masses have cooled. If there is mixing of the surface ocean with deeper and cooler water, however, surface waters cannot warm up. This was the situation before 2.7 million years ago, evidence for which comes from the high accumulation rates of diatom skeletal remains at the study site, implying vigorous diatom productivity in the overlying sea surface and therefore a continuous supply of nutrients from deeper waters.

At 2.7 million years ago, the abundance of diatom remains plummets, suggesting a



Figure 2 Icy evidence in the Northern Hemisphere: a present-day ice-sheet on the Svalbard Islands.

decrease in the nutrient availability, like that brought about by the sea surface being cut off from the deeper ocean, at least on a seasonal basis. At the same time, the reduction in vertical mixing allows the sea surface to warm. Thus the development of a seasonally layered, or stratified, surface ocean 2.7 million years ago, which was probably a regional response to the large-scale climatic changes at this time¹², allowed late summer/autumn warming of the sea surface and provided a moisture source for ice growth.

Haug *et al.*⁸ test the interpretations of the geochemical records with a suite of numerical computer-model experiments. The simulated ocean is 'stratified' and 'destratified' to determine whether this mechanism can account for the geochemically derived changes in temperature. And it can. The stratified model state produces more extreme seasons and a larger North American ice sheet than does the destratified model.

This is an exemplary study. The individual climate indicators may not have withstood the uncertainties and assumptions that limit each of them, but put together by Haug *et al.* they tell a cogent story of the origin of the ice ages. ■

Katharina Billups is at the College of Marine Studies, University of Delaware, 700 Pilottown Road, Lewes, Delaware 19958, USA.

e-mail: kbillups@udel.edu

1. Zachos, J. *et al.* *Science* **292**, 686–693 (2001).
2. Raymo, M. E. & Ruddiman, W. F. *Nature* **359**, 117–122 (1992).
3. Pearson, P. N. & Palmer, M. R. *Nature* **406**, 695–699 (2000).
4. Milankovitch, M. *Serb. Akad. Beogr. Spec. Publ.* **132** (1941).
5. Haug, G. H. & Tiedemann, R. *Nature* **393**, 673–676 (1998).
6. Driscoll, N. W. & Haug, G. H. *Science* **282**, 436–438 (1998).
7. Ravelo, A. C. *et al.* *Nature* **429**, 263–267 (2004).
8. Haug, G. H. *et al.* *Nature* **433**, 821–825 (2005).
9. Wefer, G., Berger, W. H., Bijma, J. & Fischer, G. in *Use of Proxies in Paleoclimatology* (eds Fischer, G. & Wefer, G.) 1–68 (Springer, Berlin, 1999).
10. Brandiss, M. E., O'Neil, J. R., Edlund, M. B. & Stoermer, E. F. *Geochim. Cosmochim. Acta* **62**, 1119–1125 (1998).
11. Emiliani, C. *J. Geol.* **63**, 538–578 (1955).
12. Haug, G. H. *et al.* *Nature* **401**, 779–782 (1999).

coiled tube forming the auditory part of the inner ear (Fig. 1a). The sensory hair cells reside in a thin strip of tissue, the organ of Corti, that is wedged between two membranes of the cochlea. The vibrations cause a shearing motion between these two membranes, which bends the hair bundles. Like a rolled-up piano, one end of the organ of Corti vibrates best at low frequencies and the other at high frequencies. In normal ears, the vibration is boosted and sharpened for soft sounds by what has become known as the cochlear amplifier³.

Twenty years ago, a remarkable discovery by Brownell and colleagues⁴ seemed to show what the cochlear amplifier is made of: they found that electrical stimulation of the outer hair cells made them lengthen and shorten their cell bodies. The idea is that, *in vivo*, the electricity produced by bending the hair bundles would drive this lengthening and shortening, or electromotility, as fast as sound could vibrate the bundles. The strategic position of the outer hair cells would locally boost the vibration of the organ of Corti, and in this way stimulate, by fluid coupling, the bundles on inner hair cells.

At a molecular level, this mechanism is thought to rely on a motor protein called prestin, named from the musical term for a very fast tempo. The basolateral membranes of the outer hair cells are packed with this protein⁵, which changes shape as fast as you can change the voltage across the membrane, over a range of frequencies up to at least 100 kHz (ref. 6). But there is a snag: although prestin is quick, it is not clear whether the transmembrane voltage *in vivo* changes much over the period of the sound wave, at sound frequencies greater than a few kilohertz. This is because the receptor potential due to the transducer currents is severely attenuated at higher frequencies by the electrical impedance of the cell⁷.

An alternative source of force that is not voltage-dependent may thus be needed to power the cochlear amplifier. Kennedy and colleagues¹ report large forces generated by the hair bundles of rat outer hair cells *in vitro*, when stimulated by a flexible glass fibre. You would expect the tip of the fibre to move less when attached to the bundle than when it is freely moving in the fluid. So there must have been disbelief in the lab when, in some cases, the fibre moved further when coupled to the hair bundles, implying that a force in the bundle drags the fibre along, instead of the other way round.

This force—an order of magnitude larger than the force that is a necessary by-product of opening the ion channels through which the transducer current enters hair cells⁸—is not there at the moment the hair bundle is moved by the fibre, but develops within a fraction of a millisecond. Its time course is closely coupled to that over which the transducer current adapts to a steady

Hearing

Aid from hair force

Corné Kros

Mammals hear with exquisite sensitivity and precision over a huge range of frequencies; tiny amplifiers in the inner ear make this possible. New results challenge current thinking on how these amplifiers work.

Our ability to hear relies on cells in the inner ear called hair cells — named after the bundle of 100 or so hair-like projections that protrudes from their upper surfaces. Sound bends the hair bundles, causing small electrical ('transducer') currents to flow, which in turn makes the hair cells signal the reception of sound to the brain. In mammals, the silent majority of the hair cells (the outer hair cells) do not talk to the brain, instead helping the inner hair cells — the true sensory receptors — to do so with more clarity than they could

achieve by themselves. But how is this done?

For two decades scientists have sought the answer in the extraordinary ability of the outer hair cells to change their length rapidly when stimulated. Now, however, Kennedy, Crawford and Fettiplace¹ (page 880 of this issue) and Chan and Hudspeth² (in *Nature Neuroscience*) present provocative evidence that the main component of the elusive 'cochlear amplifier' may instead reside in the hair bundles of the outer hair cells.

Sound waves that reach the ear lead to vibrations inside the cochlea—a fluid-filled,

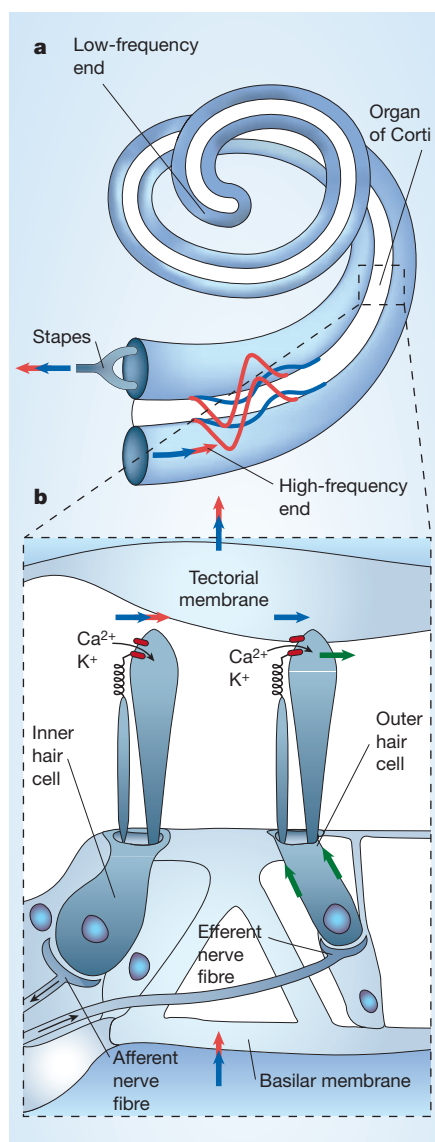


Figure 1 Cochlear amplification. a, The cochlea. Sound hitting the eardrum is transformed into vibrations of the middle ear bones. The smallest, the stapes, couples the vibration into the cochlea. The organ of Corti is found along the length of the cochlea, sandwiched between two membranes; which part vibrates depends on the frequency. The cochlear amplifier increases and sharpens the vibrations in response to soft sounds (blue curve, passive vibration; red, cochlear amplifier working). b, Cross-section through part of the organ of Corti. Shearing between tectorial and basilar membranes opens channels in the bundles of outer hair cells, causing K^+ and Ca^{2+} ions to flow in. The cochlear amplifier in these cells (green arrows, force generator in the hair bundle and/or prestin protein in the cell membrane) enhances the vibrations in different parts of the organ of Corti (blue arrows, no amplification; red arrows, amplifier active). This boosts the membranes' motion and, via fluid coupling, the hair bundles of the inner hair cells. Afferent nerve fibres contact the inner hair cells and signal sound reception to the brain; efferent nerve fibres allow the brain to control outer hair cells. Arrows indicate motion in the excitatory direction.

displacement. This calcium-dependent adaptation is extremely fast in mammalian outer hair cells⁹, and may help them to respond best to sound frequencies appropriate for their position along the cochlear spiral.

This work shows that the hair bundles of outer hair cells contain a fast force generator that does not suffer from the speed limitations of being voltage dependent. But could it really provide amplification *in vivo*? This is where Chan and Hudspeth's study² comes in. Their experimental approach was to take a turn of the gerbil cochlea and put it in an *in vitro* environment that painstakingly recreated the normal, *in vivo*, situation. It was thus possible to stimulate the organ of Corti with sound and to record the motion of the hair bundles of the inner hair cells.

The authors observed evidence for amplification of the bundle motion at low sound intensities. The amplification disappeared when the transducer current was pharmacologically blocked. This by itself does not prove that the amplification is in the hair bundle: no transducer current also means no receptor potential and hence no prestin-driven electromotility. However, when Chan and Hudspeth then prevented most of the transducer current (normally carried by potassium ions), just leaving the small fraction that is carried by calcium ions, the receptor potentials of the outer hair cells would have been much attenuated (although these were not measured). But the amplification in the motion of the inner-hair-cell bundles remained. The conclusion is that calcium-dependent force generation by the outer-hair-cell bundles may be sufficient to drive the cochlear amplifier (Fig. 1b), without the need for electromotility.

So what of prestin? Genetically engineered mice lacking this protein do not have

normal cochlear amplification, to which it must therefore make some contribution¹⁰. Prestin might work at a more leisurely pace, keeping the hair bundles at their most sensitive position and mediating the brain's ability to turn the noise down, through activation of the nerve fibres that signal to the outer hair cells. Perhaps it should be renamed 'andantin'.

These first demonstrations of force generation by the hair bundles of outer hair cells¹ and their effects on cochlear amplification² were obtained with cells from the low-frequency part of the cochlea, and it needs to be shown that they also apply at the speed limit of mammalian hearing. Other evidence is still missing, too. Do the forces measured by Kennedy and colleagues¹ disappear when transduction is blocked? Why do Chan and Hudspeth's results² not quite match the *in vivo* performance, even at low frequencies? What are needed now are experiments and models that lead to a precise, quantitative understanding of the balance of power between prestin and hair-bundle forces in cochlear amplification.

Corné Kros is in the School of Life Sciences, University of Sussex, Falmer, Brighton BN1 9QG, UK.

e-mail: c.j.kros@sussex.ac.uk

1. Kennedy, H. J., Crawford, A. C. & Fettiplace, R. *Nature* **433**, 880–883 (2005).
2. Chan, D. K. & Hudspeth, A. J. *Nature Neurosci.* **8**, 149–155 (2005).
3. Dallos, P., Popper, A. N. & Fay, R. R. (eds) *The Cochlea* (Springer, New York, 1996).
4. Brownell, W. E., Bader, C. R., Bertrand, D. & de Ribaupierre, Y. *Science* **227**, 194–196 (1985).
5. Zheng, J. *et al.* *Nature* **405**, 149–155 (2000).
6. Frank, G., Hemmert, W. & Gummer, A. W. *Proc. Natl Acad. Sci. USA* **96**, 4420–4425 (1999).
7. Santos-Sacchi, J. J. *Neurosci.* **12**, 1906–1916 (1992).
8. van Netten, S. M. & Kros, C. J. *Proc. R. Soc. Lond. B* **267**, 1915–1923 (2000).
9. Kennedy, H. J., Evans, M. G., Crawford, A. C. & Fettiplace, R. *Nature Neurosci.* **6**, 832–836 (2003).
10. Liberman, M. C. *et al.* *Nature* **419**, 300–304 (2002).

Photonics

Expect more delays

Joe T. Mok and Benjamin J. Eggleton

Slow light research has been a fast-moving topic in recent years, with potential applications from quantum computing to telecommunications. Techniques are now emerging that can slow down light in optical fibres.

Light travels at a speed c of 300 million metres per second in a vacuum, but can be slowed down to cycling speed (around 17 metres per second)¹ or can even be brought to a halt^{2,3}, when the right medium is used. Although we may not see commercial applications appearing immediately, there is a clear potential for making practical use of slow light. Recently, Song *et al.*⁴ reported a technique to delay light in optical fibres, aiming at applications in fibre-optic communication networks.

An example of a device where slow light

could be particularly useful is an all-optical router. Routers are used in communication systems to direct information from one point to another. Whereas today's routers function by first converting the information sent in optical form into electronic form, all-optical routers use all-optical switching schemes, eliminating the optical–electronic–optical conversion and are therefore inherently fast. The realization of an all-optical router requires an optical buffer — a component that functions as temporary optical storage, to effectively synchronize data packets. This

Box 1 SBS and slow light

Stimulated Brillouin scattering (SBS), which Song *et al.*⁴ use to slow down light pulses, is a coherent scattering process in an optical fibre where a forward-travelling wave ('pump') generates a refractive index grating in the form of an acoustic wave. This grating in turn reflects the pump in the backward direction through Bragg scattering, shifting it down in frequency by an amount equal to the frequency of the acoustic wave (since energy is conserved). It manifests as loss of the forward-travelling pump wave, but as gain of a 'probe' wave that is injected at the other end of the fibre and is travelling in the backwards direction, if the probe has the appropriate downshifted frequency, relative to the pump.

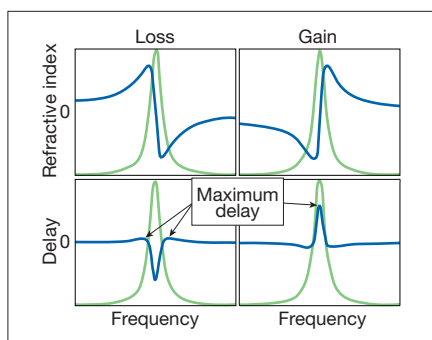
Like any other resonant systems, the SBS gain and loss spectra $\alpha(\omega)$ and its

refractive index $n(\omega)$ are uniquely correlated through the Kramers–Kronig relationship,

$$n(\omega) = 1 + \frac{c}{\pi} \int_0^\infty \frac{\alpha(\omega')}{\omega'^2 - \omega^2} d\omega' \quad (1)$$

where a positive α denotes loss, c is the speed of light and ω is angular frequency. However, the factor by which a pulse of light (which is in essence a group of light waves, all with slightly different frequencies) is slowed down is the group index n_g , which is related to n through

$$n_g = n + \omega \frac{dn}{d\omega} \quad (2)$$



where the speed of the light pulse is then given by c/n_g . Pulse delay is proportional to n_g .

The figure shows plots of equations (1) and (2) with Lorentzian loss and gain resonances (green lines) and their associated refractive indices (a) and group indices, or delay (b). In particular, a large delay occurs when there is a large positive change in refractive index. For the case of loss, this occurs at the sides of the resonance peak. For the case of gain, however, maximum delay coincides with the peak of the resonance.

Song *et al.* take advantage of this last effect and demonstrate that a light pulse can simultaneously be delayed and experience a significant gain (undergo amplification) with the SBS mechanism. **JTM & BJE**

comparison, a tuning mechanism based on optical intensity, as demonstrated by Song *et al.*, can be more mechanically stable and easily implemented. In this case, the tuning speed is only limited by how quickly the pump intensity can be adjusted.

Another advantage of the work, as well as of the previous delay-tunable devices^{6,7}, is that they are based on optical fibres, and are therefore compatible with present-day telecommunications networks. An important difference between the three is their bandwidth, which varies from tens of megahertz to tens of gigahertz. A larger bandwidth implies that the device is capable of slowing shorter pulses, and thus is more desirable in, for example, communication networks operating at high bit-rates. However, systems with larger bandwidth generally have a smaller range of tunable delay⁸. A solution to this trade-off involves using a longer device to achieve larger delay without sacrificing bandwidth, but at the expense of compactness and elegance. The challenge ahead is thus to achieve a large delay tunability over a large bandwidth with short device length.

Many view slow light as an important aspect of our future capability to process and transport information optically. This is exemplified by the US\$6.5 million funding from the US Defense Advanced Research Projects Agency into a programme named 'Applications of Slow Light in Optical Fibers', in addition to another \$18 million into two research programmes related to all-optical routers. From the application point of view, the relevant performance measure for a slow light device is not the absolute delay achieved by the device but its bandwidth, range of tunability and compactness. Higher bandwidth is desirable for ultrafast high-bit-rate applications, whereas delay tunability is essential in time-critical applications where packets of information can be released from the buffer at will. The various demonstrations of fibre-based, delay-tunable devices are a step closer to real-life applications of slow light. ■

Joe T. Mok and Benjamin J. Eggleton are at the ARC Centre of Excellence for Ultrahigh-bandwidth Devices for Optical Systems (CUDOS), School of Physics, University of Sydney, Sydney, NSW 2006, Australia.

e-mail: j.mok@physics.usyd.edu.au; egg@physics.usyd.edu.au

- Hau, L. V., Harris, S. F., Dutton, Z. & Behroozi, C. H. *Nature* **397**, 594–598 (1999).
- Liu, C., Dutton, Z., Behroozi, C. H. & Hau, L. V. *Nature* **409**, 490–493 (2001).
- Turukhin, A. V. *et al. Phys. Rev. Lett.* **88**, 0236021–0236024 (2002).
- Song, K. Y., Herráez, M. G. & Thévenaz, L. *Opt. Express* **13**, 82–88 (2005).
- Bigelow, M. S., Lepeshkin, N. N. & Boyd, R. W. *Science* **301**, 200–202 (2003).
- Heebner, J. E., Wong, V., Schweinsberg, A., Boyd, R. W. & Jackson, D. J. *IEEE J. Quant. Elect.* **40**, 726–730 (2004).
- Eggleton, B. J., de Sterke, C. M. & Slusher, R. E. *J. Opt. Soc. Am. B* **16**, 587–599 (1999).
- Lenz, G., Eggleton, B. J., Madsen, C. K. & Slusher, R. E. *IEEE J. Quant. Elect.* **37**, 525–532 (2001).

is what a slow light device would be able to perform. Another potential application of slow light is in quantum computing², which promises to be many times more powerful than electronic computing. In this context, a slow light device can act as the optical equivalent of today's electronic random access memory (more commonly known as RAM).

The key to slowing down light lies in using a carefully chosen spectral portion of an optical resonance. Light does not travel uninhibited through a medium that is optically resonant, but interacts with the medium so that its propagation characteristics, and in particular its speed, can be changed. Research groups around the world are investigating a variety of materials that exhibit resonant behaviour to obtain slow light — using quantum interference effects^{1,5} or building microstructures such as ring resonators⁶ and photonic crystals⁷. Song *et al.*⁴ make use of a mechanism called stimulated Brillouin scattering (SBS) (see Box 1).

In previous work where slow light was demonstrated using a resonance effect, the delay of an optical pulse was always associated with loss, with the undesirable consequence that the optical signal was attenuated. What differentiates the work of Song *et al.* is that they demonstrate a resonant gain

effect — the frequency of maximum gain coincides with that of maximum delay in their SBS approach (Box 1, Fig. b). So, when the right frequency is chosen, a light pulse is amplified while being slowed down.

The authors send intense light (the 'pump') from one end of an optical fibre, and at the same time inject from the other end a weaker light pulse (the 'probe'). The pump produces, through the SBS mechanism, a resonance in the fibre that is in tune with the frequency spectrum of the probe. This changes the propagation properties of the probe pulse and leads to both its amplification and delay (Box 1). Song *et al.* demonstrate that a probe pulse of 100 nanoseconds can be slowed down by 30 nanoseconds, while experiencing 1,000-fold amplification.

An interesting aspect of Song *et al.*'s work is that they demonstrate an optically variable delay; the precise speed to which the probe pulse is slowed down can be tuned by the intensity of the pump wave, as the authors show both theoretically and experimentally. Although this is the first demonstration of variable pulse delay by optical means, other delay-tuning mechanisms for slow light devices already exist. They include tuning of strain in an optical fibre that contains a photonic crystal structure⁷ and wavelength-tuning in a fibre ring resonator⁶. In

Immunology

Guide for a cell-fate decision

Ellen A. Robey

How does an immature cell know how to develop into a specialized one? A fortunate observation has revealed one of the cues that guide precursor immune cells to their ultimate fate.

During the development of multicellular organisms, precursor cells mature into specialized cell types, such as muscle cells or blood cells. This cellular differentiation must be carefully orchestrated to generate the correct numbers and types of cells at the right time and place. Unravelling the regulatory circuitry involved is a major goal of developmental biologists. In a striking illustration of Louis Pasteur's famous statement, "Chance favours only the prepared mind", a fortuitous mutation in a mouse has led Kappes and colleagues to a breakthrough in understanding how immune-cell precursors adopt their appropriate cell fate (He *et al.*¹, page 826 of this issue).

In the mammalian thymus, precursor cells called thymocytes give rise to two types of mature immune cell: CD4 helper T cells, which alert the immune system when the body is invaded by a pathogen and coordinate an inflammatory response; and CD8 killer T cells, which destroy cells that have been invaded. Each T cell has receptors for a specific antigen on its surface. The gene that encodes these receptors is rearranged during the cell's development to produce one of a

huge potential variety of receptor genes. The range of receptors consequently produced by the entire population of mature T cells allows the immune system to recognize virtually any pathogen. These T cells probe other cells they come across, and respond when they find 'foreign' peptides that might signal the presence of a pathogen. The foreign peptides become bound to molecules of the major histocompatibility complex (MHC) and are displayed on the surface of invaded cells and other immune cells. The MHC comes in two varieties: class I, which is recognized by the CD8 killer T cells, and class II, which is recognized by the CD4 helper T cells. The CD8 and CD4 proteins aid the interaction between the MHC and the T-cell antigen receptor (Fig. 1).

In the thymus, immature T cells express both CD4 and CD8. The cells test whether their newly formed antigen receptors can bind to MHC complexes on thymic epithelial cells that display 'self' peptides, derived from the body's own proteins. Most of the thymocytes fail the test and die. Remarkably, binding to the peptide–MHC complex not only leads to the survival of thymocytes, but also directs them to the appropriate lineage:

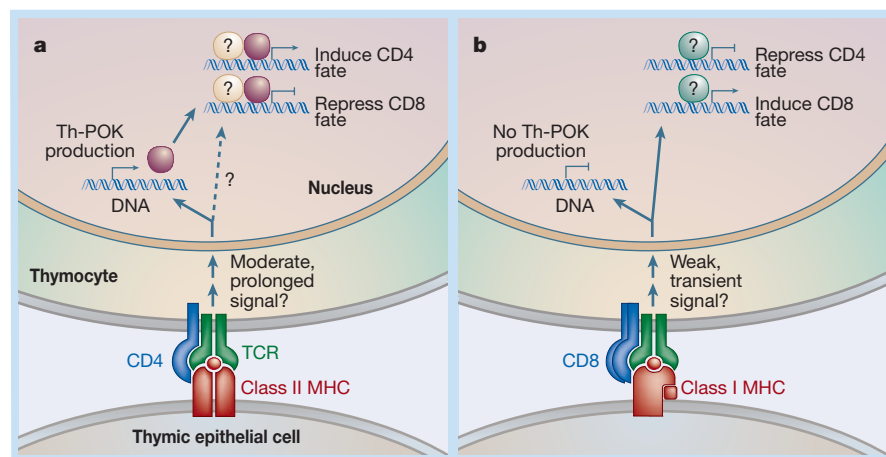


Figure 1 The anatomy of a cell-fate decision. In the thymus, immune-cell precursors called thymocytes receive positive selection signals when their T-cell receptors (TCRs) recognize 'self' peptides (red circles) bound to molecules of the major histocompatibility complex (MHC) displayed on thymic epithelial cells. This interaction is helped by the CD4 or CD8 proteins. **a**, The thymocyte recognizes class II MHC, and receives a moderate, prolonged signal. This leads to production of Th-POK, a gene-regulatory protein that in turn controls other genes (including repression of the CD8 gene), causing the cell to adopt a CD4 fate. TCR signalling may also work independently of Th-POK (dashed line) to promote the differentiation or survival of CD4 cells. **b**, The thymocyte recognizes class I MHC, and receives a weak, transient TCR signal that fails to increase production of Th-POK. TCR-induced events in the absence of Th-POK lead to changes in gene expression (including repression of the CD4 gene), causing the cell to adopt the CD8 fate.

recognition of class I MHC leads to development of a CD8 T cell, and recognition of class II MHC leads to development of a CD4 T cell (Fig. 1).

The work by Kappes and colleagues¹ defines the molecular trail from the thymocyte–MHC interaction to the change in gene expression that guides the cell to its ultimate fate. While screening mice bred from two lines carrying different genetically engineered mutations, the authors noted² that some of the progeny had an unusual characteristic that was not the result of either mutation — a complete absence of mature CD4 T cells. A spontaneous mutation, which they termed 'helper deficient', or HD, was responsible for the defect. Further analysis revealed that HD did not simply cause loss of CD4 cells, but rather produced a 'cell-fate switch' by which thymocytes that would normally have matured into CD4 T cells instead gave rise to CD8 T cells³.

Although there are a number of engineered mutations that can also redirect the CD4 versus CD8 fate decision (reviewed in refs 4, 5), most of these mutations lead to proteins that have abnormal functions (rather than causing disruption of a normal function), and thus they may not reflect the true functions of the unmutated genes. Furthermore, most of these engineered mutations cause only a fraction of thymocytes to make the wrong cell-fate choice. HD is the only mutation known that not only is a loss-of-function mutation, but also causes a wholesale conversion of thymocytes from the CD4 to the CD8 lineage, and so it seems to hold the key to understanding this cell-fate decision.

Seven years after the original description of the HD mutation, Kappes and colleagues¹ have now identified the gene affected by it. It encodes a transcription factor called Th-POK (for T-helper-inducing POZ/ Krüppel factor). Transcription factors regulate gene activity, often in a tissue-specific manner. The discovery, therefore, that the HD mutation changes an amino acid of a transcription factor made perfect sense. The authors also show that, in the thymus, Th-POK is found specifically in mature CD4 T cells, and not in CD8 T cells. Moreover, enforced expression of the Th-POK gene in thymocytes transforms cells that would have been CD8 into CD4 cells, the opposite effect to the HD mutation. Thus, Th-POK is necessary and sufficient to induce the gene-expression programme for a CD4 cell fate and to suppress the CD8 programme.

Other transcription factors have been implicated in T-cell development, including GATA3, which is required for CD4 T-cell maturation^{6,7}, and Runx3, which silences the gene encoding CD4 in CD8 T cells⁸. However, mutation of their corresponding genes does not cause a cell-fate switch, but merely impairs the development of one cell



100 YEARS AGO

A Grass-snake which the writer had in his possession for eighteen months has just died. A fact which seems worthy of a note is the length of time during which this snake fasted. The last time the snake fed was June 11, 1904, the meal consisting of a small frog. From that time until the date of its death, February 2, it took no food, although constantly offered it. The animal thus existed for close on eight months without food. During the whole of this time it appeared in good health, and was, at times, most animated. No approach to hibernation was observed, and only for a little more than a week before its death did the snake seem out of health. The body was not unduly thin. From *Nature* 23 February 1905.

50 YEARS AGO

The organization for the fishing and conservation of the *Cœlacanth*s of the Comoro Islands created by the *Institute de Recherche Scientifique de Madagascar*... reports a new success: on November 12 last a further *Latimeria* was captured at Anjouan. This brings the total since 1938 to eight and is the finest yet, as regards both size and state of preservation, and by far the most interesting because it is the first near-adult female specimen which has come into our hands as well as the first of these precious fishes which anyone has been able to observe alive... Throughout the night — which the delighted population of Mutsamudu passed in singing and dancing to celebrate the capture — the *Cœlacanth* was watched over with admirable care by the chef de circonscription, taking turns with his adjoint, M. Solère. It seemed, although quite bewildered at the sequel to its ascent to the surface, to be taking the situation very well, swimming slowly by curious rotating movements of its pectoral fins, while the second dorsal and anal, likewise very mobile, served together with the tail as a rudder. After daybreak it became apparent that the light, and above all the sun itself, was upsetting the animal very much [and] the fish began to show more and more obvious signs of distress... At 14.45 hr. it was still swimming feebly; but at 15.30 hr. it had its belly in the air and only the fins and gill-covers were making agonized movements.

From *Nature* 26 February 1955.

type. So Th-POK seems to be at the top of a gene-regulation hierarchy that controls T-cell fate.

Many tantalizing questions remain. What controls expression of the Th-POK gene in developing thymocytes? The observation that Th-POK synthesis is increased in thymocyte precursors that recognize class II MHC, but not class I MHC, implies that the gene's expression is regulated by engagement of class II MHC. This may be controlled by differences in T-cell receptor signalling on recognition of class I compared with class II MHC, as predicted by a popular model for CD4 versus CD8 lineage commitment (reviewed in refs 4, 5). Further analysis of the regulation of Th-POK by T-cell-receptor signalling will be needed to test this idea.

What genes are regulated directly by Th-POK? The gene-expression programmes of CD4 and CD8 cells differ in terms of the production of GATA3, perforin (a T-cell effector protein), and CD4 and CD8 themselves. However, it is not yet clear whether these differences in gene expression are direct or indirect consequences of Th-POK activity.

Finally, self-reinforcing feedback loops are often used to 'lock in' cell-fate decisions, and there are hints that Th-POK is part of such a regulatory circuit. HD mice expressing class II MHC, but not class I MHC, show increased synthesis of messenger RNA for Th-POK in thymocyte precursor cells. However, the redirected CD8 T cells that ultimately develop in these mice lack Th-POK. Whether the requirement for Th-POK

function to maintain Th-POK gene expression involves auto-regulation of the gene or involves other components of the gene regulatory hierarchy remains to be seen.

The quest to understand the CD4 versus CD8 lineage decision has focused mostly on the question of whether MHC recognition instructs the cell-fate decision (the instructive model), or whether cells choose their fate randomly and are then tested for the presence of the appropriate CD4 or CD8 molecule (the stochastic or selection model)^{9,10}. These models, although useful, assumed that T-cell fate determination occurs as a single discrete step, but it has become clear that it is a multi-step process involving feedback and reinforcement. The identification of Th-POK as a key in the T-cell fate decision should open the way to deeper mechanistic insights into how cells are guided to their destiny. ■

Ellen A. Robey is in the Department of Molecular and Cell Biology, 471 Life Sciences Addition, University of California, Berkeley, California 94720, USA.
e-mail: erobey@uclink.berkeley.edu

1. He, X. *et al.* *Nature* **433**, 826–833 (2005).
2. Dave, V. P., Allman, D., Keefe, R., Hardy, R. R. & Kappes, D. J. *Proc. Natl Acad. Sci. USA* **95**, 8187–8192 (1998).
3. Keefe, R., Dave, V., Allman, D., Wiest, D. & Kappes, D. J. *Science* **286**, 1149–1153 (1999).
4. Basson, M. A. & Zamoyska, R. *Immunol. Today* **21**, 509–514 (2000).
5. Germain, R. N. *Nature Rev. Immunol.* **2**, 309–322 (2002).
6. Hernandez-Hoyos, G., Anderson, M. K., Wang, C., Rothenberg, E. V. & Alberola-Ila, J. *Immunity* **19**, 83–94 (2003).
7. Pai, S. Y. *et al.* *Immunity* **19**, 863–875 (2003).
8. Taniuchi, I. *et al.* *Cell* **111**, 621–633 (2002).
9. Janeway, C. A. *Nature* **335**, 208–210 (1988).
10. Robey, E. A. *et al.* *Cell* **64**, 99–107 (1991).

Planetary science

Being there

Kevin Zahnle

The protoplanets that collided to make the Earth may themselves have had atmospheres and oceans. Venus has vastly more argon and neon than Earth: fossil evidence, perhaps, of protoplanetary atmospheres?

Noble gases are the flotsam of the Solar System. They seem simple: they shun chemistry, they are difficult to freeze, and they accumulate in atmospheres. We see them as passive tracers of our cosmogonic theories, especially theories that address the origin and evolution of planetary volatiles. Perhaps it is a measure of our theories that every newly probed atmosphere has surprised us. Yet the temptation to read profound meaning into noble-gas abundances remains strong. On page 842 of this issue, Genda and Abe¹ ask whether atmospheres can survive a series of giant impacts, such as the collision with Earth that formed our Moon. The answers are ambiguous, with argon agreeable but neon cryptic.

Current fashion posits four stages in the growth of an Earth: coagulation of grains into boulders; gathering of the boulders into aggregates of kilometre dimensions; runaway accretion of those aggregates into Moon-sized protoplanets; and giant collisions between the protoplanets to make planets². The first three stages are thought to have taken no more than a million years in total, whereas the fourth played out over tens of millions of years or more.

Given how quickly stages 1–3 occurred, there is a fair likelihood that significant nebular gas was still present when the protoplanets emerged from runaway accretion. If the nebula was cold enough, and the protoplanets large enough, they would gravitationally capture significant primary atmospheres.

When the nebula disappeared, the proto-planets would retain some of these atmospheres and thereby acquire noble gases from the solar nebula.

Genda and Abe¹ use a ruthlessly simplified model to describe how giant impacts interact with atmospheres². They accept as a matter of course that much of the stricken hemisphere is lost, driven off by hydrodynamic jetting or the tangential momentum of a glancing blow. Their concern is with the hemisphere that is not hit directly.

Here the impact announces itself when the bow shock erupts through the surface. Indeed, the entire crust launches into space at high velocity, typically several kilometres per second. The crust drives a shock wave into the atmosphere that accelerates as it rises into thinner air. At some height the air reaches escape velocity, but escape is efficient only if the ground velocity approaches or exceeds the escape velocity^{3,4}. Such high ground velocities are difficult to achieve if the planet itself survives; the highest ground motions achieved in successful Moon-forming impact simulations (a relatively gentle event) are typically less than half the escape velocity⁵. Genda and Abe therefore argue that most of a planet's atmosphere, including its primordial noble gases, is retained in giant impacts.

An ocean changes everything. When the crust hits the ocean, the ocean is driven outward at an even higher velocity and it is also vaporized. The resulting explosion of supercharged steam accelerates most of the overlying atmosphere to escape velocity and beyond (Fig. 2 in the supplementary information¹). Genda and Abe therefore argue that, with an ocean, most of the atmosphere (and most of the noble gas) is lost in giant impacts. Because they were there, the noble gases provide the best test of their hypothesis.

The authors take the minority view that bound water was common to all the building blocks of Earth and Venus, so that from the start the protoplanets each had their own oceans and atmospheres, and independent lives as infant Earths. A more popular view is that Earth and Venus became wet while still accreting, but only because they were struck by cold, wet protoplanets ejected from what is now the asteroid belt⁶; with modest revision, however, Genda and Abe's hypothesis could apply in this context as well.

To distinguish between Earth and Venus, Genda and Abe invoke the runaway greenhouse effect⁷. On worlds nearer the Sun, water evaporates into a thick atmosphere, whereas on more distant worlds it condenses as oceans and ice sheets. With oceans, the building blocks of Earth lost much more of their primary atmospheres than did the building blocks of Venus. In this way, Genda and Abe's model can account for Venus

having 70 times more argon than Earth.

Of course there are caveats. Genda and Abe stress low-velocity collisions that merge the two protoplanets. A broader assessment of impact geometries and velocities suggests that most giant impacts are bounces, rather than mergers, in which both bodies emerge slightly smaller (and much hotter) amidst a spray of silicates⁸. In a high-speed bouncing impact between unequal planets, the smaller one could easily lose both its crust and its atmosphere.

But the big problem with gravitational capture of nebular noble gases is the abundance of neon. The neon:argon ratio on Venus, Earth, Mars and meteorites is invariably about 1% of what it is in the Sun⁸. Explaining this requires both a heroic theory of selective neon escape⁹ and some luck to make the neon:argon ratio always come out the same. It may work better to start from some other source of noble gases that discriminates against neon — say, extremely low-temperature condensates that quantitatively trap argon¹⁰ (neon freezes only if hydrogen freezes), or noble gas implantation by the solar wind into unaccreted grains and boulders^{11–13}. If you merely wish to bury the argon, there is no limit to how much.

Human immunodeficiency virus

Refolding the envelope

Peter D. Kwong

HIV has evolved to avoid neutralization by human antibodies. New atomic-level details reveal that such evasion involves substantial refolding of its exterior glycoprotein.

A defining feature of HIV is the ability to thrive for a decade or more despite sustained and vigorous immune attack. How does the virus remain able to infect host cells at the same time as evading immune surveillance? The solution resides largely in the molecular trickery of a single viral protein, gp120 — the component of the 'spikes' on the viral surface that binds to receptors on host cells.

Tantalizing new details of this trickery are revealed on page 834 of this issue¹, where Chen and colleagues report the structure of the unliganded (receptor-free) form of gp120. Comparison with the previously determined structure² of gp120 in complex with one of its cell-surface receptors reveals that receptor binding induces roughly half of the protein to refold. Like a trick jigsaw puzzle that can form two entirely different pictures, the refolding preserves pieces of gp120's secondary structure, but reshuffles their location and relative orientation. In the viral spike, this reshuffling moves pieces of protein by more than 40 Å, rivalling the

The new work offers a fresh look at planetary accretion. The argument that oceans speed the loss of atmospheres is more broadly applicable than the authors imply. It need not be restricted to atmospheres of solar composition acquired gravitationally; and it need not be restricted to water oceans or even liquid oceans. It applies generally during or before the giant-impact stage of accretion, and it could apply to satellites of the outer Solar System as easily as to Earth.

Kevin Zahnle is at the NASA Ames Research Center, Space Science Division, MS 245-3, Moffett Field, California 94035-1000, USA.

e-mail: kevin.j.zahnle@nasa.gov

1. Genda, H. & Abe, Y. *Nature* **433**, 842–844 (2005).
2. Lissauer, J. J. *Annu. Rev. Astron. Astrophys.* **31**, 129–174 (1993).
3. Chen, G. Q. & Ahrens, T. J. *Phys. Earth Planet. Int.* **100**, 21–26 (1997).
4. Genda, H. & Abe, Y. *Icarus* **164**, 149–162 (2003).
5. Agnor, C. & Asphaug, E. *Astrophys. J. Lett.* **613**, 157–160 (2004).
6. Morbidelli, A. et al. *Meteorit. Planet. Sci.* **35**, 1309–1320 (2000).
7. Nakajima, S., Hayashi, Y. & Abe, Y. *J. Atmos. Sci.* **49**, 2256–2266 (1992).
8. Pepin, R. O. *Icarus* **92**, 2–79 (1991).
9. Zahnle, K. J., Kasting, J. F. & Pollack, J. B. *Icarus* **84**, 502–527 (1990).
10. Owen, T. et al. *Nature* **402**, 269–270 (1999).
11. Wetherill, G. *Icarus* **46**, 70–80 (1981).
12. Sasaki, S. *Icarus* **91**, 29–38 (1991).
13. Ballentine, C., Marty, B., Lollar, B. S. & Cassidy, M. *Nature* **433**, 33–38 (2005).

largest changes ever observed for a single protein domain.

Soon after HIV was identified as the aetiological agent of AIDS, it became clear that although infection prompts the human body to produce a large number of HIV-reactive antibodies, they are mostly ineffective at neutralizing the virus³. The machinery behind HIV's remarkable evasion involves a protective membrane around the virus and an ordered, two-receptor mechanism of entry into cells. The membrane is derived from a previous host cell and forms the outer surface of HIV — its envelope. The only viral proteins that protrude through this membrane are the envelope glycoproteins, gp120 and gp41: three copies of each make up a viral spike.

The two-receptor mechanism involves the virus first binding to the CD4 glycoprotein on the surface of a potential host cell. This binding induces changes in gp120 that permit it to interact with a second cell-surface receptor, or co-receptor^{4,5}. This mechanism is not essential for viral propagation,

as variants that do not require CD4 for entry (CD4-independent isolates) can be obtained *in vitro*⁶. Rather, the mechanism appears to be necessary for immune evasion, as alterations in gp120 that lead to CD4 independence are invariably linked to increased sensitivity to neutralization by antibody.

The upshot of all this is that, in HIV isolates that resist neutralization, gp120 can exist in two quite different states. Before it interacts with the cell surface, it is highly resistant to antibody binding — hidden by the overlapping defences of steric occlusion, conformational masking and glycan (sugar) shielding^{7–9}. After binding to CD4, gp120 is primed for interaction with other proteins, including the co-receptor. In this 'CD4-induced' state, it would in theory be sensitive to antibody binding. But the sensitizing trigger occurs at the interface between the virus and the host cell, and the close proximity to the cell membrane sterically inhibits antibody binding¹⁰.

The structure of gp120 before CD4 binding has been sought for almost 20 years, but the machinery that prevents antibody binding also inhibits the lattice interactions necessary for X-ray analysis. In a technical *tour de force*, however, Chen and colleagues¹ have now been able to grow crystals of an unliganded gp120 core, build an atomic-level structure, model the expected orientation of the unliganded gp120 in the viral spike, and outline a pathway for how CD4 binding triggers large conformational rearrangements.

The results show that the unliganded gp120 attaches to the rest of the viral spike through a highly hydrated inner domain, suggesting that it would retain considerable flexibility in the assembled spike. Sequence-variable regions from neighbouring gp120s point towards each other; by interlocking to restrict gp120 movements, they provide a means for the variable regions to control conformational masking⁸.

N-linked glycans cover much of the surface of gp120 that would be expected to form the spike's outer face, and one can envisage them forming a shield to prevent antibody binding⁹. Much of the glycan has an ordered structure, implying that glycans at high density may display a limited range of conformations. Structural elements needed to bind the gp120 co-receptor — elements that, when assembled together, would in theory be sensitive to antibodies — are spatially separated in the unliganded structure, and only assemble upon CD4 binding (Fig. 1). All of these details lay the groundwork for understanding how gp120 in the viral spike can evade antibodies.

Several issues remain. First, the relationship between the monomeric form of gp120 studied by Chen and colleagues and the oligomeric form in the viral spike is unclear. Does oligomeric gp120 fold into a different

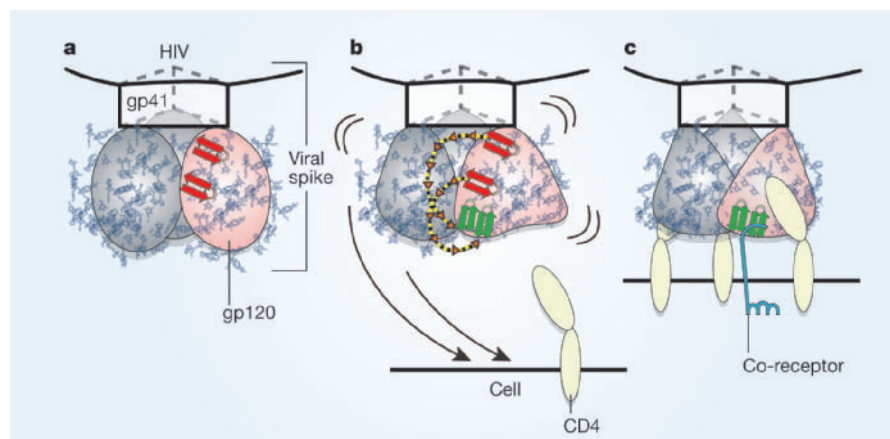


Figure 1 Molecular tricks of HIV. **a**, Before HIV attaches to the surface of a host cell, its surface spikes (one of which is shown here) are accessible to antibodies, but protected by numerous viral defences. These include N-linked glycans (blue), which cloak the three gp120 glycoproteins (pink and grey ovals) and three gp41s (triangular plate) that make up a spike. **b**, Chen and colleagues¹ find that attachment to the cell's CD4 protein induces roughly half of the gp120 glycoprotein molecule to refold. This refolding is highlighted for two β -ribbons (red), which rearrange (green) to form part of a site that allows binding to a second receptor, the co-receptor. **c**, The crowded virus–cell interface protects the assembled co-receptor-binding site from antibodies.

conformation? Possibly so: secondary-structure analysis¹¹ strongly predicts that portions of gp120 that are involved in co-receptor binding should have α -helical structure, but they adopt mostly β -strand configurations in the unliganded¹ and CD4-bound² structures. Second, the unliganded structure does not contain regions of the protein that govern protein–protein interactions in the spike.

So it remains to be seen how spike interactions fix gp120 in such a way as to allow CD4-induced rearrangements while preventing antibody binding. A complete understanding of antibody resistance may require snapshots not only of each conformation of gp120, but of the entire spike.

The classic 'before' and 'after' images of the fusion peptide of influenza virus being thrown 100 Å towards the host-cell membrane shattered the staid notion of proteins having a single unique manner of folding¹². Such extensive refolding seemed, however, to be a special facet of the process of fusing viral and host-cell membranes during viral entry — a process requiring large overall molecular movements. The newly revealed 'before' image of gp120 now shows that HIV deploys extensive refolding not only for fusion but also to evade neutralization by antibody.

Paradoxically, however, the refolding that protects HIV from antibodies may also be a source of vulnerability. One consequence of folding into more than one conformation is that dual packing restraints may leave empty pockets or cavities. Both the unliganded and the CD4-bound gp120 structures contain unusual pockets. One such pocket in the unliganded structure is, Chen and colleagues find, the binding site for a recently discovered class of viral-entry inhibitors¹³. Although the authors' unliganded gp120

crystals are not suitable for structure-based drug design, tinkering with lattice contacts and with this site of inhibitor binding may permit suitable complex crystals to form.

In terms of vaccine design, the structure of unliganded gp120 reveals the envelope at its potentially most vulnerable. This is the state before attachment — before occlusion at the cell interface — where HIV should be most accessible to antibody binding. But this is also the state in which the virus uses its full arsenal of tricks to evade antibody neutralization.

The details revealed by Chen and colleagues will enable each of these tricks to be probed by the increasingly sophisticated tools of protein design, to fix or expose otherwise transient or occluded sites of antibody binding. Will such modified envelopes — based on the unliganded structure — be capable of eliciting broadly neutralizing antibodies? In addition to altering the way we view receptor binding, this structure¹ opens the envelope on these and other experiments.

Peter D. Kwong is at the Vaccine Research Center, National Institutes of Health, 40 Convent Drive, Bethesda, Maryland 20892, USA.

e-mail: pdkwong@nih.gov

- Chen, B. *et al.* *Nature* **433**, 834–841 (2005).
- Kwong, P. D. *et al.* *Nature* **393**, 648–659 (1998).
- Weiss, R. A. *et al.* *Nature* **316**, 69–72 (1985).
- Feng, Y. *et al.* *Science* **272**, 872–877 (1996).
- Wu, L. *et al.* *Nature* **384**, 179–183 (1996).
- Kolchinsky, P. *et al.* *J. Virol.* **73**, 8120–8126 (1999).
- Wyatt, R. *et al.* *Nature* **393**, 705–711 (1998).
- Kwong, P. D. *et al.* *Nature* **420**, 678–682 (2002).
- Wei, X. *et al.* *Nature* **422**, 307–312 (2003).
- Labrijn, A. F. *et al.* *J. Virol.* **77**, 10557–10565 (2003).
- Rost, B., Yachdav, G. & Liu, J. *Nucleic Acids Res.* **32**, W321–W326 (2004).
- Bullough, P. A., Hughson, F. M., Skehel, J. J. & Wiley, D. C. *Nature* **371**, 37–43 (1994).
- Lin, P. F. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 11013–11018 (2003).

Obituary

Eduard Kellenberger (1920–2004)

When asked, at the 1964 Swiss national exhibition, if one “could be a good Swiss citizen and get up late”, Eduard Kellenberger answered “yes”. He was not referring to himself. He was an invariably early riser, one who has left an enduring legacy to Swiss — and European — science. His contributions to electron microscopy and the genetics of bacterial viruses (phages), as well as his institutional achievements, made him the earliest representative of molecular biology in Switzerland.

In many ways, Kellenberger's career was typical of the now-vanishing generation that founded European molecular biology. He studied physics at the Swiss Federal Institute of Technology in Zurich under Paul Scherrer, the most prominent atomic physicist in Switzerland. In 1945, he joined the laboratory of Jean Weigle at the University of Geneva to work on developing a Swiss-made industrial electron microscope. To illustrate the usefulness of such an instrument for biomedical research, Kellenberger attempted to provide micrographs for the sales catalogue. At that time, however, preparation methods were so crude that biological samples were either destroyed by the electron beam or so distorted as to be meaningless. Focusing on bacteria and phages, Kellenberger eventually succeeded in 1958, together with the chemist Antoinette Rytter, in developing the ‘RK-method’, which went on to become a standard protocol, and their paper a citation classic.

The scientific priorities of the atomic age brought many young men and women into physics, but they also pushed some physicists out of the discipline. Weigle, for example, who did not wish to work in the large teams of ‘big science’, left for the California Institute of Technology in 1948, where he became a phage geneticist, a close collaborator of Max Delbrück, the spiritual father of the rapidly growing Phage Group. By returning every year to his former laboratory, then led by Kellenberger, Weigle created a crucial transatlantic bridge, making the Geneva group part of a handful of European researchers working on phage genetics.

The scientific reputation of Kellenberger's laboratory resulted from a unique combination of electron microscopy and phage genetics. Unlike many microscopists, or physicists more generally, Kellenberger not only sought instrumental virtuosity but had a deep



Pioneer in electron microscopy

desire to understand the biological problems at stake, such as the function of genes involved in phage assembly. As a result, in the early 1960s, Kellenberger's laboratory became a central source of phage and bacterial micrographs for molecular biology textbooks. These micrographs also made a strong impression on a broader public, including university and political authorities. Crucially, they gave these key administrators a visual appreciation of the promise of biology at the molecular level.

Kellenberger cherished freedom of thought. In his laboratory meetings every idea was welcome, even the apparently naive. His intellectual generosity and respect for the work of others created a relaxed atmosphere that encouraged individual initiative. In 1959, for example, he hired the physicist Werner Arber to investigate the mutagenic effects of X-rays. Two years later, Arber had still not unpacked the X-ray apparatus as he had found himself enmeshed instead in the study of ‘host-controlled variation’. Rather than pressuring him to move back on track, Kellenberger supported Arber through what appeared to be a highly unfashionable project — wisely so, as Arber's work eventually led to the discovery of restriction enzymes and to a Nobel prize in 1978.

Kellenberger did much to secure molecular biology a place on the science policy agendas, both nationally and at the European level. Until the late 1950s, the term ‘molecular biology’ enjoyed little currency and was largely undefined. As early as 1959, however, Kellenberger began

the planning of an Institute of Molecular Biology at the University of Geneva. In addition to his biophysics laboratory, centred on electron microscopy and phage genetics, Kellenberger conceived of a biochemistry laboratory to be headed by Alfred Tissières, then working with James Watson at Harvard. In his mind, structural, biochemical and genetic approaches were a necessary combination for molecular biological research, a vision shared by an increasing number of scientists around the world. In 1963, the institute was created in Geneva, along the lines of Kellenberger's plan.

That same year, Kellenberger took part in the founding of the European Molecular Biology Organization (EMBO) and sat on its first council. EMBO wished to foster molecular biology in Europe, even though it began as nothing more than a club of scientists, without funding. Kellenberger arranged for EMBO to become a legal organization under Swiss law and then directly contacted the Swiss minister of foreign affairs with a letter entitled “What is life?”. This must have had an impact, because the minister engaged the Swiss government in taking the delicate diplomatic initiative of bringing European states to support EMBO financially. Once an intergovernmental agreement was reached, EMBO began planning a biology equivalent of CERN, the European particle physics facility — leading to the opening of the European Molecular Biology Laboratory in Heidelberg.

After having spent 25 years at the University of Geneva, interrupted only by occasional sabbaticals in the United States, Kellenberger decided to take up another challenge and, in 1970, helped found a new interdisciplinary research institution at the University of Basel: the Biozentrum. At that time, molecular biology was becoming increasingly established in European institutions, thanks to those, such as Kellenberger, who early on had understood its potential.

Kellenberger did not defend science at all cost, however, but was driven by a strong humanist ideal that led him to question the place of atomic energy, and later biotechnology, in society, long before it was fashionable to do so. Until his death on 13 December 2004, Kellenberger was a vigorous proponent of socially responsible globalization, driven by the desire to make knowledge serve the common good. For all this we are profoundly grateful to him.

Bruno J. Strasser and Jacques Dubochet are at the University of Lausanne, Dorigny, CH-1015 Lausanne, Switzerland.
e-mail: bruno.strasser@unil.ch

Entomology

Unlocking the sex secrets of cockroaches

Science **307**, 1104–1106 (2005)

Humans have notched up a victory in the long-running battle with cockroaches. Satoshi Nojima *et al.* have purified and identified the sex pheromone of the female German cockroach (*Blattella germanica*), more than a decade after the gland that produces it was discovered. The finding could pave the way for more effective cockroach traps.

The pheromone had proved resistant to isolation by standard gas chromatography — in which components evaporate out of a mix one by one — because it fell apart in the heat. Using a new, lower-temperature version of the procedure, Nojima *et al.* isolated the contents of the gland and then used a cockroach antenna to single out the active pheromone. Finally they used nuclear magnetic resonance analysis to determine the structure of the compound.

The pheromone is gentsyl quinone isovalerate — or, the authors propose, ‘blattellaquinone’. When the team baited traps with it in the lab and at a roach-infested pig farm, male roaches came running. Given the compound’s structure, Nojima *et al.* suggest that it may have been used for defence earlier in the insect’s evolutionary history, although the quantities found in females are now too small to be useful for this purpose.

Emma Marris

Neurobiology

Beyond a supporting role

Cell **120**, 421–433 (2005)

Biologists once viewed astrocytes as passive supports for neurons. But as actors in the bigger picture, these cells are not simply extras. That point has been made to striking effect in a study by Karen S. Christopherson and colleagues — their research provides evidence that astrocytes produce proteins called thrombospondins that encourage the growth of new neuronal synapses.

Christopherson *et al.* added a purified form of these proteins to rat brain cells and witnessed synapse growth strikingly similar to that induced by astrocyte secretions. Exactly how the proteins work remains unclear, but in the future the authors hope to pinpoint the neuronal receptor to which they bind.

The presence of thrombospondins in young, developing brains suggests that these molecules have an important role in the formation of synapses. Moreover, their levels are decreased in adults, perhaps helping to explain why adult brains are less ‘plastic’ than younger ones.

Roxanne Khamis



Pattern formation

Square rules

Phys. Rev. Lett. **94**, 054503 (2005)

The archetype for the division of two-dimensional space into cells has long been the honeycomb. Its hexagonal symmetry is echoed in natural phenomena as diverse as bubble rafts and geological formations such as the Giant’s Causeway in County Antrim, Northern Ireland. For biological cells, the paradigm tends to be threefold junctions of cell walls that meet at 120°.

But many naturally arising patterns have a different geometry, exemplified by the filigree of cracks in a ceramic glaze (pictured). Here the cells are predominantly four-sided, and cracks tend to meet at right angles.

Bohn *et al.* show that this pattern can be understood by considering the sequential, hierarchical manner in which the cracks form. Long, space-crossing rifts come first, and then the space in between is filled with ever-finer, stress-relieving fractures that do not alter the existing fissures.

The researchers use geometric principles such as Euler’s theorem (relating the number of network vertices and edges to the number of cells they define) to show why, in this hierarchical formation, the cells will on average be four-sided but with six neighbours. Intriguingly, this kind of pattern is also seen in the street networks of old cities (before urban planning skewed the topology), where a few major routes radiating from the centre were gradually supplemented by ever-narrower roads, lanes and alleys in between.

Philip Ball

Atmospheric science

Plant food from pollution

J. Geophys. Res. doi:10.1029/2004JD005082 (2005)

Iron is an essential nutrient for phytoplankton, the tiny aquatic plants that carry out almost half of all photosynthesis on Earth.

Dust storms in northern China and Mongolia carry iron from the soil of the Gobi desert to the northern Pacific Ocean. But the iron in desert dust is in a mineral form that has low solubility in seawater and so is not readily available to phytoplankton. Nicholas Meskhidze and colleagues have found that sulphur dioxide pollution from industrial plants in China can acidify the dust, which converts iron to a more soluble form.

The team tracked two dust storms from the Gobi desert that passed over Beijing and then the Pacific in 2001. Using satellite measurements, they saw an increase in phytoplankton growth after a storm in March, but, surprisingly, no increase after a larger storm in April. They attribute this difference to the fact that the dust in the April storm contained much more calcium carbonate, which neutralizes SO₂ and therefore limits the process of making iron more soluble.

Natural sources of SO₂, such as volcanoes, may also boost phytoplankton growth. This mechanism could be vital for fertilizing the oceans, the authors add, increasing the uptake of CO₂ during photosynthesis.

Mark Peplow

Materials chemistry

Magnetic sensors

Chem. Mater. doi:10.1021/cm0486971 (2005)

Magneto-optic memories are now commercially available. But C. Michael Elliott *et al.* report that they have developed a material that changes its response to light in the presence of weak magnetic fields, and propose that magnetism and light can be coupled to give an optical read-out of data from a magnetic storage medium.

In their molecular ‘triad’ — comprising an electron-donating group, an electron acceptor and a light-absorbing chromophore (a dye) — charge is transferred between the donor and acceptor when the chromophore is illuminated. The length of time the acceptor remains in this charged state is altered by a weak magnetic field, providing a possible ‘detection’ signal. However, this usually only happens when triad species are in solution, and only solid materials can make magneto-optical storage practical.

Elliott *et al.* get around the problem by encapsulating droplets of an aqueous solution of their triad in a polymeric matrix. They first sequester the droplets inside reverse micelles of a surfactant dispersed in a polymerizable organic liquid. Crosslinking the solvent molecules then produces a hard, transparent polymer in which the charge-transfer dye retains its optical properties.

Laser light is used to excite the charged state, and a modest magnetic field induces a measurable shift in its lifetime.

Philip Ball

A synthetic enamel for rapid tooth repair

Seamless fixing of an early caries lesion can be achieved without prior excavation.

The conventional treatment of dental caries involves mechanical removal of the affected part and filling of the hole with a resin or metal alloy^{1–4}. But this method is not ideal for tiny early lesions^{5,6} because a disproportionate amount of healthy tooth must be removed to make the alloy or resin stick. Here we describe a dental paste of synthetic enamel that rapidly and seamlessly repairs early caries lesions by nanocrystalline growth, with minimal wastage of the natural enamel.

The human tooth is protected by enamel of 1–2 mm thickness that is composed of hydroxyapatite ($\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2$) crystals. In early caries lesions, acid-forming bacteria cause microscopic damage to the enamel, creating cavities that are less than 50 μm deep. Such cavities cannot be repaired by simple setting of restorative materials because these do not adhere perfectly to the enamel owing to differences in chemical composition and crystal structure.

We prepared a white crystalline paste of modified hydroxyapatite, which chemically and structurally resembles natural enamel, and used it to repair an early caries lesion in a lower premolar tooth (for methods, see supplementary information).

The affected site was sealed off within 15 min. Examination of the microstructure of the restoration using transmission electron microscopy (TEM) reveals no obvious structural gap at the interface between the regrown layer and the enamel region (Fig. 1a). The regrown layer contains elongated crystals (100–400 nm long and 20–80 nm wide) that have grown across the interface and are regularly orientated to the tooth surface. This shows that the paste has properly integrated with the tooth enamel.

An atomic-resolution TEM image of a grown crystal (Fig. 1b) reveals that it has two lattice periodicities that are consistent with the inter-lattice distances for the *c* and *a* directions of a hydroxyapatite crystal (0.688 and 0.817 nm, respectively). From these results, combined with those from X-ray photoelectron spectroscopy analysis (data not shown), we conclude that the hydroxyapatite crystals in the regrown layer are oriented with their (0001) face parallel to the tooth surface⁷. The regrown layer contains about 1% by atoms of fluoride ions and has a calcium-to-phosphorus molar ratio of 1.58 ± 0.03 . It also has

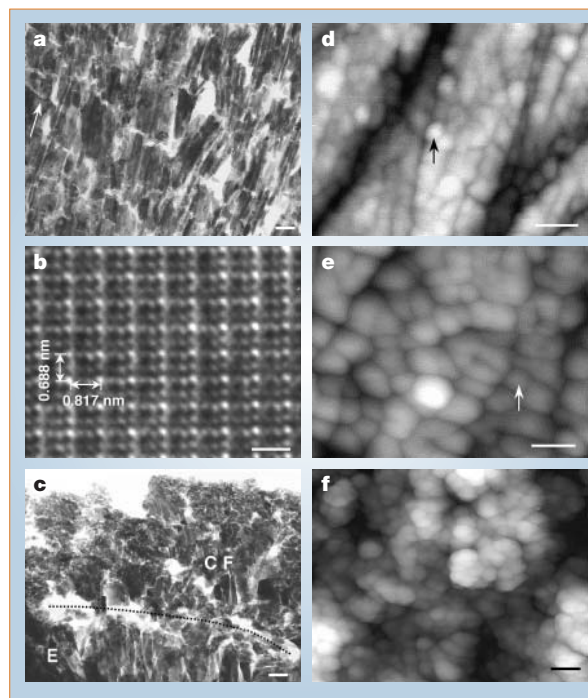


Figure 1 Repair of an early caries lesion using a synthetic enamel paste. **a–c**, Transmission electron micrographs, and **d–f**, atomic-force microscopy images of tooth repair. **a**, Image of the interface between the regrown layer (upper region) and the enamel (lower region); the arrow indicates the direction of the tooth surface. **b**, Atomic image of a grown crystal of synthetic enamel. **c**, Tooth treated with acidic phosphate fluoride; dotted line indicates the interface between the calcium fluoride layer (CF) and enamel (E). **d**, Original enamel surface; arrow, hydroxyapatite crystal. **e**, Surface after 3 min of repair time; arrow, newly grown hydroxyapatite crystal. **f**, Surface after completion of repair (15 min). Scale bars: **a, c, f**, 100 nm; **b**, 1 nm; **d, e**, 50 nm.

high durability and acid tolerance (see supplementary information).

For comparison, we repaired a similar lesion with acidic phosphate fluoride solution, an alternative treatment for early caries lesions that does not necessitate the removal of healthy tooth enamel, and examined the restoration by TEM (Fig. 1c). The image shows the presence of a calcium fluoride layer of inhomogeneous thickness^{8,9} (and less than 1 μm thick) covering the enamel, and a clear gap at the interface (Fig. 1c, dotted line).

Time-lapse atomic-force microscopy indicates that the hydroxyapatite crystals of the original tooth enamel (Fig. 1d) initially dissolved slightly during the repair, but quickly grew again because the paste was acting as a source of crystals. This dissolution and regrowth occurs as a result of the strong acidity ($\text{pH} < 2$) of the mother solution and paste. The process creates a continuous, nanometre-scale structure that extends from the enamel to the regrown layer by epitaxial growth of crystals.

The newly grown crystals of hydroxyapatite cover the whole surface in a densely

packed array after 3 min (Fig. 1e), and are three-dimensionally stacked after 15 min (Fig. 1f). The acidic conditions probably contribute to the fast growth of highly crystalline hydroxyapatite by dissociating the calcium phosphate clusters into calcium and phosphate ions^{10,11}; the clusters would otherwise slow growth rates and cause low crystallinity.

We have shown that our synthetic material can reconstruct enamel without prior excavation, in a process that not only repairs early caries lesions but can also help to prevent their reoccurrence by strengthening the natural enamel. In the clinic, the paste should not come into contact with the gums, where its acidity and its high concentration of hydrogen peroxide could cause inflammation (though materials with similarly adverse properties are already used on patients¹²).

Kazuo Yamagishi*, **Kazuo Onuma†**, **Takashi Suzuki‡**, **Fumio Okada‡**, **Junji Tagami§**, **Masayuki Otsuki§**, **Pisol Senawangse§**

*FAP Dental Institute, 3-2-1, Kakinokizaka, 502 Meguro-ku, Tokyo 152-0022, Japan
e-mail: FZT02705@nifty.com

†National Institute of Advanced Industrial Science and Technology, Institute for Human Science and Biomedical Engineering, Central 6, 1-1-1 Higashi,

Tsukuba, Ibaraki 305-8566, Japan

‡Department of Applied Chemistry and Biotechnology, Faculty of Engineering, Yamanashi University, 43-11, Takeda, Kofu, Yamanashi 400-8511, Japan

§Department of Restorative Science, Graduate School, Tokyo Medical and Dental University, 5-45, Yushima 1-chome, Bunkyo-ku, Tokyo 113-8549, Japan

1. Raskin, A., Michotte-Theall, B., Vreven, J. & Wilson, N. H. F. *J. Dent.* **27**, 13–19 (1999).

2. Wilson, N. H. F. & Mjor, I. A. *J. Dent.* **28**, 15–21 (2000).

3. Carvalho, R. M., Pereira, J. C., Yoshiyama, M. & Pashley, D. H. *Oper. Dent.* **21**, 17–24 (1996).

4. Hilton, T. J. *Am. J. Dent.* **15**, 198–210 (2002).

5. Frank, R. M. & Brendel, A. *Archs Oral Biol.* **11**, 883–912 (1966).

6. Johnson, N. W. *Caries Res.* **1**, 356–369 (1967).

7. Elliot, J. C. (ed) *Structure and Chemistry of the Apatites and Other Calcium Orthophosphates* (Elsevier, Amsterdam, 1994).

8. Gerould, H. *J. Dent. Res.* **24**, 223–233 (1945).

9. Duschner, H., Gotz, H. & Ogaard, B. *Eur. J. Oral Sci.* **105**, 466–472 (1997).

10. Onuma, K. & Ito, A. *Chem. Mater.* **10**, 3346–3351 (1998).

11. Banfield, J. F., Welch, S. A., Zhang, H., Ebert, T. T. & Penn, R. L. *Science* **289**, 751–754 (2000).

12. Yamagishi, K. & Suzuki, T. *J. Esthetic Dent.* **7**, 78–80 (1995).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Ecology

A niche for cyanobacteria containing chlorophyll *d*

The cyanobacterium known as *Acaryochloris marina* is a unique phototroph that uses chlorophyll *d* as its principal light-harvesting pigment instead of chlorophyll *a*, the form commonly found in plants, algae and other cyanobacteria; this means that it depends on far-red light for photosynthesis. Here we demonstrate photosynthetic activity in *Acaryochloris*-like phototrophs that live underneath minute coral-reef invertebrates (didemnid ascidians) in a shaded niche enriched in near-infrared light. This discovery clarifies how these cyanobacteria are able to thrive as free-living organisms in their natural habitat.

Acaryochloris marina was first isolated from extracts of didemnid ascidians^{1,2} and was presumed to be a symbiont, like the cyanobacterium *Prochloron* sp., which contains chlorophyll *a* and *b*, and is found inside didemnids³. *Acaryochloris marina* has been found on red algae⁴ and a free-living *Acaryochloris*-like organism has been discovered in a turbid saline lake⁵. This indicates that cyanobacteria containing chlorophyll *d* may be fairly widespread, yet little is known about their habitat and ecology.

In a microphotometric survey of the didemnid ascidians *Lissoclinum patella*,

Trididemnum paracyclops, *Diplosoma similis* and *Diplosoma virens*, we investigated the occurrence and distribution of cells containing chlorophyll *d* (for methods, see supplementary information). *Prochloron*, and some unicellular cyanobacteria containing chlorophyll *a* and phycobiliproteins, colonized internal cavities of the didemnids, but we found no evidence of chlorophyll *d* in the ascidians. However, biofilms growing on the underside of the didemnids contained clusters of pale, greenish-yellow *Acaryochloris*-like morphotypes: these had spectral absorption and fluorescence features that were characteristic of chlorophyll *d* (Fig. 1a, b; and see supplementary information).

We cultured these *Acaryochloris*-like cells from the biofilm. Sequence analysis of the isolate (results not shown) indicates that the genes that encode cells' light-harvesting protein (*pcbC*) and 16S ribosomal RNA correspond to those of *A. marina*: there is 100% identity with the 302-base-pair polymerase-chain-reaction fragment and 99% identity with the 392-base-pair fragment, respectively (NCBI database⁶).

Fibre-optic microprobe measurements⁷ in *D. virens* showed intense attenuation of visible light. Far-red light penetrated more efficiently through the ascidian tissue, and was enhanced relative to the incident light owing to light-trapping effects^{3,7}. Under all ascidians, visible light was strongly depleted but there was 10–20 times more far-red light,

providing an ideal niche for cyanobacteria containing chlorophyll *d*, which absorbs maximally at 700–720 nm (Fig. 1b).

We used variable chlorophyll fluorescence imaging to assess photosynthesis of the cyanobacteria containing chlorophyll *d* in their natural habitat. Maximal quantum yields of photosystem II (PSII) were 0.77 and 0.59 in zones comprising *Prochloron* and *Acaryochloris*-like cells, respectively (Fig. 1c). Maximal PSII quantum yields of 0.67–0.80 have been reported for *A. marina* cultures⁸.

As expected, the quantum yield of PSII decreased with increasing irradiance (Fig. 1d). Surprisingly, the *Acaryochloris*-like cells were, like *Prochloron*, able to sustain high photosynthetic activity at strong light intensity (Fig. 1e). A similar light adaptation is also evident in *A. marina*⁹.

It is an apparent paradox that *Acaryochloris*-like cells thrive in extreme shade but show features of adaptation to strong light. This unusual photoacclimation reflects the fact that they live in an environment rich in near-infrared light and that chlorophyll *d* is the main light-harvesting pigment that drives both photosystems I and II under these conditions^{7,10}.

We conclude that *Acaryochloris*-like cyanobacteria grow in biofilms beneath didemnid ascidians, where far-red is enhanced over visible light and is used for oxygenic photosynthesis. This explains the occurrence of epiphytic *A. marina* on the underside of red algae⁴. Cyanobacteria that contain chlorophyll *d* may thrive in other habitats with little visible light, but further microenvironmental controls may be important in defining the niche of these microorganisms.

Michael Köhl*, Min Chen†, Peter J. Ralph‡, Ulrich Schreiber§, Anthony W. D. Larkum†

*Marine Biological Laboratory, Institute of Biology, University of Copenhagen, 3000 Helsingør, Denmark
e-mail: mkühl@bi.ku.dk

†School of Biological Sciences, A08, University of Sydney, New South Wales 2006, Australia

‡Institute for Water and Environmental Resource Management, University of Technology Sydney, Gore Hill, New South Wales 2065, Australia

§Julius-von-Sachs-Institut für Biowissenschaften, Universität Würzburg, 97082 Würzburg, Germany

1. Miyashita, H. et al. *Nature* **383**, 402 (1996).
2. Miyashita, H., Ikemoto, H., Kurano, N., Miyachi, S. & Chihara, M. *J. Phycol.* **39**, 1247–1253 (2003).
3. Köhl, M. & Larkum, A. W. D. in *Symbiosis: Mechanisms and Model Systems* (ed. Seckbach, J.) 273–290 (Kluwer, Dordrecht, 2002).
4. Murakami, A., Miyashita, H., Iseki, M., Adachi, K. & Mimuro, M. *Science* **303**, 1633 (2004).
5. Miller, S. R. et al. *Proc. Natl Acad. Sci. USA* **102**, 850–855 (2005).
6. Chen, M., Hiller, R. G., Howe, C. J. & Larkum, A. W. D. *Mol. Biol. Evol.* **22**, 21–28 (2005).
7. Köhl, M. & Fenchel, T. *Microb. Ecol.* **40**, 94–103 (2000).
8. Schiller, H., Senger, H., Miyashita, H., Miyachi, S. & Dau, H. *Fed. Eur. Biochem. Soc. Lett.* **410**, 433–436 (1997).
9. Miyashita, H. et al. *Plant Cell Physiol.* **38**, 274–281 (1997).
10. Boichenko, V. A., Klimov, V. V., Miyashita, H. & Miyachi, S. *Photosynth. Res.* **65**, 269–277 (2000).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

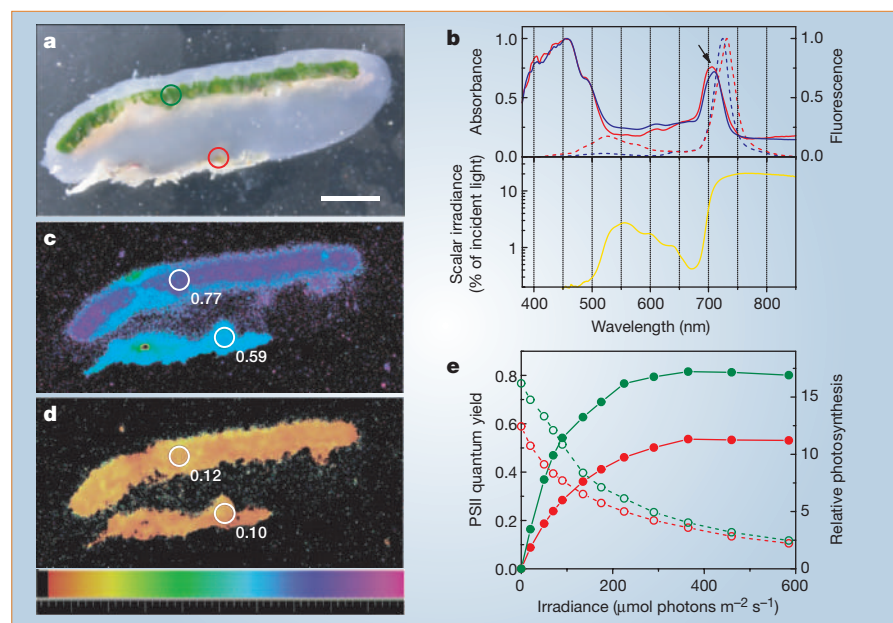


Figure 1 Distribution, spectral characteristics and photosynthesis of cells containing chlorophyll *d* that are associated with the didemnid ascidian *Diplosoma virens*. **a**, Vertical section through *D. virens*, showing the green cells of symbiotic cyanobacterium *Prochloron* sp. inside cavities, and a biofilm (white) patch of *Acaryochloris*-like cells growing on the underside of the ascidian (scale bar, 2 mm). **b**, Top, spectral absorbance (solid lines) and ultraviolet-excited fluorescence (dashed lines) of cells from the biofilm shown in **a** (red curves) and cells from an *A. marina* culture (blue curves). Arrow, absorption maximum of chlorophyll *d*. Data were normalized to the maximal absorbance and fluorescence, respectively. Bottom, spectral irradiance measured below *D. virens* after the biofilm had been removed, expressed as a percentage of downwelling irradiance at the tissue surface. **c**, **d**, Images as in **a**, but showing the maximal photosystem-II (PSII) quantum yield of the dark-adapted section (**c**) and the effective PSII quantum yield at an irradiance of 585 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ (**d**). Both variables were scaled to the same colour gradient (0–1). **e**, PSII quantum yield (dashed lines) and relative rates of photosynthesis (solid lines) as a function of irradiance in *Prochloron* symbionts (green) and *Acaryochloris*-like cells (red), taken from areas circled in **a**.

North Pacific seasonality and the glaciation of North America 2.7 million years ago

Gerald H. Haug¹, Andrey Ganopolski², Daniel M. Sigman³, Antoni Rosell-Mele⁴, George E. A. Swann⁵, Ralf Tiedemann⁶, Samuel L. Jaccard⁷, Jörg Bollmann⁷, Mark A. Maslin⁵, Melanie J. Leng⁸ & Geoffrey Eglinton⁹

¹Geoforschungszentrum Potsdam (GFZ), and ²Potsdam Institute for Climate Impact Research (PIK), 14473 Potsdam, Germany

³Department of Geosciences, Princeton University, Princeton, New Jersey 08544, USA

⁴ICREA and ICTA, Autonomous University of Barcelona, 08193 Bellaterra, Catalonia, Spain

⁵Environmental Change Research Centre, Department of Geography, University College London, London, WC1H 0AP, UK

⁶IFM-Geomar, 24148 Kiel, Germany

⁷Department of Earth Sciences, ETH Zürich, 8092 Zürich, Switzerland

⁸NERC Isotope Geosciences Laboratory, British Geological Survey, Keyworth, Nottingham NG12 5GG, UK

⁹Biogeochemistry Centre, University of Bristol, Bristol BS8 1TS, UK

In the context of gradual Cenozoic cooling, the timing of the onset of significant Northern Hemisphere glaciation 2.7 million years ago is consistent with Milankovitch's orbital theory, which posited that ice sheets grow when polar summertime insolation and temperature are low. However, the role of moisture supply in the initiation of large Northern Hemisphere ice sheets has remained unclear. The subarctic Pacific Ocean represents a significant source of water vapour to boreal North America, but it has been largely overlooked in efforts to explain Northern Hemisphere glaciation. Here we present alkenone unsaturation ratios and diatom oxygen isotope ratios from a sediment core in the western subarctic Pacific Ocean, indicating that 2.7 million years ago late-summer sea surface temperatures in this ocean region rose in response to an increase in stratification. At the same time, winter sea surface temperatures cooled, winter floating ice became more abundant and global climate descended into glacial conditions. We suggest that the observed summer warming extended into the autumn, providing water vapour to northern North America, where it precipitated and accumulated as snow, and thus allowed the initiation of Northern Hemisphere glaciation.

To initiate and sustain the large Northern Hemisphere ice sheets of the Plio-Pleistocene ice ages, two requirements are broadly recognized. First, the more polar continental areas must be sufficiently cold for precipitation to fall as snow rather than rain and for snow and ice to survive the warm summer melting season. Second, adequate moisture must be introduced to high northern latitudes to promote the accumulation of glacial ice. In attempts to explain the initiation of the major Northern Hemisphere glaciation 2.7 million years (Myr) ago, much attention has been given to the temperature requirements of continental glaciation. The time interval between 4.5 and 3.1 Myr was dominated by a pronounced long-term minimum in the amplitude of the 41-kyr cycle in the obliquity of the Earth's rotation¹, which would have failed to produce particularly cold Northern Hemisphere summers—the key requirement posited by Milankovitch for the onset of Northern Hemisphere glaciation. During this time interval, there may have been several aborted shifts toward glaciation, for example, between 4.1–3.9 Myr and 3.5–3.3 Myr (ref. 2; Fig. 1). During the late Pliocene and early Pleistocene, a high amplitude in the obliquity cycle resulted in periods of low tilt angle, which, in turn, would have yielded periods with cold summers in the Northern Hemisphere. Thus, it has been suggested that the progressive increase in the amplitude of the obliquity cycle tipped the scale between 3.1–2.5 Myr, allowing for long-term expansion of Northern Hemisphere ice¹. In short, our long-held view of the temperature requirement of glaciation is largely consistent with the timing of the onset of Northern Hemisphere glaciation.

However, the onset of Northern Hemisphere glaciation has proved to be inconsistent with ideas regarding the water vapour requirement^{3,4}. It has been suggested that glaciation began in response to increased North Atlantic Deep Water formation and

the flow of warm Gulf Stream waters into the high-latitude North Atlantic, associated with the closure of the Panama seaway^{5,6}. However, recent studies show that this closure and associated changes in North Atlantic circulation occurred 4.6 Myr ago, well before the onset of intense Northern Hemisphere glaciation^{4,5}. Thus, it is unknown whether and how a change in water vapour supply encouraged the initiation of intense Northern Hemisphere glaciation.

Seasonality of the modern subarctic North Pacific

The modern subarctic Pacific surface is dominated by a permanent 'halocline', or salinity-driven density gradient in the upper 300 m, that reduces exchange between the surface layer and the ocean interior⁷. Seasonal changes in the temperature of the surface mixed layer are thus only minimally buffered by the heat capacity of the ocean subsurface, resulting in a sea surface temperature (SST) that has one of the largest annual ranges of any open ocean region, with winter (February) surface ocean temperatures in the subarctic Northwest Pacific of about +1 °C, late-summer (September) temperatures of +12 °C, and a seasonal thermocline during summer and autumn⁸.

This seasonal variation in the physical conditions of the subarctic Pacific leads to strong seasonality in the biological productivity of the region. Winter mixing transports nutrients from the subsurface into the euphotic zone. During spring, as the euphotic zone deepens and the mixed layer shoals, a diatom-dominated bloom begins, lasting until early summer, when most of the nutrients are consumed, silicate in particular^{9,10}. However, during late summer and autumn, when the water column is most stable, a secondary biogenic bloom typically occurs, this time dominated by coccolithophores^{11,12}. Alkenones accumulating in the sediments below

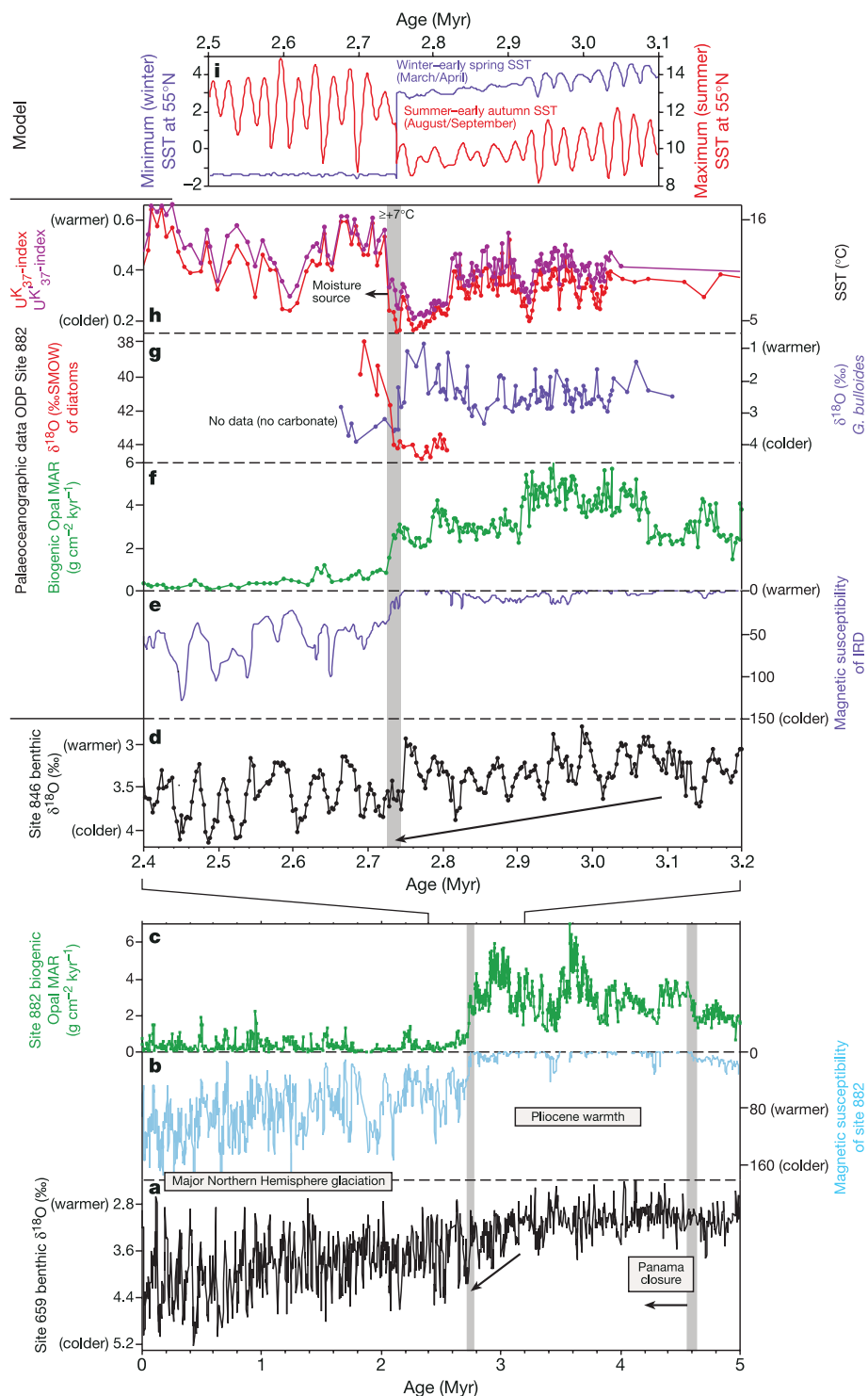


Figure 1 Palaeoceanographic data and model time series through the time interval marking the onset of major Northern Hemisphere glaciation. **a**, Increase in ice volume between 3.1 and 2.7 Myr, as indicated by benthic foraminiferal $\delta^{18}\text{O}$ from ODP Site 659, in the eastern equatorial Atlantic Ocean³⁴. **b**, IRD input to the subarctic Northwest Pacific, as indicated by the increase in magnetic susceptibility at ODP Site 882 (50° 21' N, 167° 35' E, water depth 3,244 m) at 2.73 Myr. **c**, Drop in biogenic opal mass accumulation rates (MAR) at ODP Site 882 in the subarctic Northwest Pacific. **d–h**, During the time interval 3.2 to 2.4 Myr, fluctuations in ice volume as indicated by benthic foraminiferal $\delta^{18}\text{O}$ from ODP Site 846, eastern equatorial Pacific² (**d**), IRD at ODP Site 882

(**e**), biogenic opal MARs at ODP Site 882 (**f**), $\delta^{18}\text{O}$ in planktonic foraminifera *G. bulloides* (blue), which is interpreted to reflect mainly winter/spring SST, and $\delta^{18}\text{O}$ of large diatom species *C. marginatus* and *C. radiatus* (red), which is interpreted to reflect mainly late summer/autumn SST (**g**), and U_{37}^{K} - and U_{37}^{K} -indices, which are interpreted to reflect mainly late summer/autumn SST (**h**). The range of absolute SSTs (in °C, right axis) reflect the U_{37}^{K} temperature calibration of ref. 18, which is in close agreement with that of ref. 19. **i**, CLIMBER-2 model output of minimum (blue; March/April or winter/spring) and maximum (red; August/September or summer autumn) zonally averaged Pacific SST at 55° N.

this region indicate a modern temperature of 10.1 °C, consistent with the late-summer and autumn growth of the coccolithophorids, which are among the prymnesiophytes that produce these compounds¹¹.

North Pacific changes 2.7 Myr ago

Before about 2.7 Myr ago, the accumulation of diatomaceous sediments in the subarctic Pacific was roughly four to five times greater than it is today (Fig. 1; ODP Site 882: 50° 21' N, 167° 35' E, water depth of 3,244 m). Records from both the western and eastern basins of the subarctic Pacific indicate an abrupt drop in opal accumulation rate at isotope stage G6, synchronous with the massive onset of ice-rafted debris (IRD, Fig. 1). The nearly complete consumption of silicate in the modern subarctic Pacific summer-time surface and a sedimentary N isotope change across the 2.7-Myr transition conspire to indicate that the drop in opal accumulation was associated with a drop in the supply of major nutrients from the ocean interior to the surface ocean¹³. Thus, the biogeochemical data point to the development of the subarctic Pacific halocline at 2.7 Myr, closely associated with the onset of major Northern Hemisphere glaciation¹³ (Fig. 1).

The close association of subarctic Pacific halocline formation with major Northern Hemisphere glaciation as well as the abrupt and dramatic nature of both changes suggest a positive feedback between the two. We have previously focused on how climate cooling increased the vertical stability of the North Pacific^{13,14}. This work raised atmospheric CO₂ as the possible mechanism by which polar stratification could, in turn, cause global cooling and thus participate in a positive feedback. The sediment core data and climate model output reported here provide a more direct mechanism by which the development of the subarctic Pacific halocline set the scene for ice-sheet growth in the Northern Hemisphere.

The $\delta^{18}\text{O}$ of microfossil calcite from the planktonic foraminifer *G. bulloides* increases at 2.7 Myr by approximately 2‰ (ref. 14; Fig. 1). This has been taken to indicate a drop in SST of about 5 °C, taking ice-volume variations into account¹⁵. The $\delta^{18}\text{O}$ increase occurs shortly (~3 kyr) before the drop in opal accumulation and coincides with the first step in IRD increase. Poor preservation makes such a foraminifera-derived reconstruction difficult in these nearly carbonate-free sediments, especially after the 2.7-Myr transition. Nevertheless, the sharp increase in IRD across the transition and the evidence from other regions¹⁶ confirm that the overall sense of change at 2.7 Myr was a dramatic cooling.

We have measured alkenone unsaturation ratios^{17–22} (U_{37}^K and $U_{37}^{K'}$) across the 2.7-Myr transition at ODP Site 882, to provide an additional constraint on surface temperature changes in the subarctic North Pacific. Prymnesiophytes, including the coccolithophorids, produce the long-chain alkenones that are used in this temperature reconstruction. As phytoplankton, these organisms are concentrated in the upper euphotic zone of polar waters, whereas *G. bulloides* can live at a variety of depths and also forms a gametogenic crust in the subsurface. Moreover, coccolithophorids tend to bloom in the middle to late summer in the western subarctic Pacific, after the diatom bloom^{9–12}, whereas foraminiferal production tends to follow the productivity of the entire phytoplankton pool and thus is at a maximum in the spring^{9,10}. For these reasons, significant differences should be expected between alkenone- and foraminiferal-based temperature reconstructions. Even with this expectation, the alkenone-derived temperature change is in surprising contrast to the indicators of cooling at the 2.7-Myr transition: the alkenone data indicate a warming of $\geq 7^\circ\text{C}$ across the transition (Fig. 1).

Given this unexpected result, possible sources of artefact must be considered. The alkenone content of these old, high-latitude sediments is quite low, requiring the use of a high-sensitivity gas chromatography-chemical ionization mass spectrometry (GC-CIMS) method for measurement of the degree of alkenone

unsaturation²³ (see Supplementary Information). Diagenesis and changes in light and nutrient conditions are not expected to bias the unsaturation ratios to the degree that would be required to explain the transition at 2.7 Myr (refs 24, 25). A recent concern is that alkenones can be transported laterally, associated with fine sediments, and can perhaps be remobilized from ancient sediments²⁶. This is an unlikely concern for the sediments of ODP Site 882. Sediment transport in the region is, if anything, from the North, and there is no evidence for a radical change in lateral transport at the 2.7-Myr transition. For remobilization and subsequent incorporation of alkenones from older sediments to have caused the apparent decrease in U_{37}^K at the 2.7-Myr transition, extremely old sediments would need to be eroded to produce such a 'warm' U_{37}^K signature. This assumption is not supported by the composition of the coccolith assemblage, which is dominated by coccoliths typical/indicative of the time period analysed (mainly gephyrocapsids and reticulofenestrids).

The alkenone evidence for post-2.7-Myr summertime warming is corroborated by the $\delta^{18}\text{O}$ of the silica frustules produced by large (75–150 μm) autumn-living subarctic North Pacific diatom species (*Coscinodiscus marginatus* and *C. radiatus*, see Supplementary Information). Their $\delta^{18}\text{O}$ decreases by approximately 5‰ across the 2.7-Myr transition (Fig. 1), indicating some combination of warming and freshening in the late-summertime/autumn surface. Development of the full modern halocline in the subarctic Pacific at 2.7 Myr can explain only ~1‰ of this $\delta^{18}\text{O}$ decrease⁸, whereas one would expect ice volume to have caused a global ocean $\delta^{18}\text{O}$ increase of ~0.5‰ (ref. 2). This leaves a ~4.5‰ decrease to be explained by late-summertime/autumn sea surface warming. Published coefficients for the dependence of diatom silica $\delta^{18}\text{O}$ on temperature range from 0.2‰ to 0.5‰ per °C (refs 27–29). Thus, the diatom $\delta^{18}\text{O}$ data appear to require a warming of $\geq 9^\circ\text{C}$ at 2.7 Myr, which is similar to the warming estimate from the alkenones. Although significant uncertainties remain in the use of diatom $\delta^{18}\text{O}$, they are completely different from those that apply to the alkenones. In particular, lateral transport or exhumation from older sediments are not plausible concerns for these large and extremely well-preserved diatoms (see Supplementary Information).

A link between stratification and seasonality

Despite the initially counterintuitive nature of these results, warming is completely consistent with the development of the subarctic Pacific halocline at 2.7 Myr. Because the halocline acts to reduce exchange between the surface and ocean interior, the development of the halocline at 2.7 Myr should have caused the seasonal variation of surface ocean temperature to reflect more fully the seasonal cycle in insolation and air temperature. That is, upon stratification, the seasonal variation in surface temperature should have increased toward the ~11 °C range that characterizes the modern subarctic Pacific. The U_{37}^K -index reflects the SST maximum that coincides with the late-summer/autumn coccolith bloom, and the diatoms analysed here also grow at the surface during the late summer¹⁰. By contrast, the foraminiferal calcite is biased to record the spring diatom bloom and is also strongly influenced by the temperature of the shallow subsurface through growth below the mixed layer and the formation of gametogenic crust. Consequently, the apparent paradox between the cooling at 2.7 Myr as indicated by the planktonic foraminifera $\delta^{18}\text{O}$ and the warming as indicated by the U_{37}^K -index and diatom $\delta^{18}\text{O}$ is best interpreted as an expression of the amplified seasonal contrast that should have been expected from the development of the permanent halocline at that time.

The high heat capacity of sea water causes surface waters to remain warm into the autumn and cool into the spring. Stratification of the subarctic Pacific will decrease the thermal inertia of the upper ocean and thus reduce the phase lag between land and ocean temperature. However, the amplification of the seasonal cycle

should overwhelm this effect, so that stratification will cause the subarctic Pacific surface to be significantly warmer than the land further into the autumn.

Such an enhanced temperature contrast with continental climate, which responds rapidly to seasonal insolation changes, is well suited for driving glaciation in North America. The subarctic Pacific is a dominant source of water vapour to boreal North America³⁰, and warmer SSTs in the autumn would cause a larger fraction of the water vapour delivery to occur when continental climate is cold enough for snow to accumulate. In this way, the stratification of the subarctic Pacific would allow for adequate water vapour supply to feed glaciers when global climate cooling would otherwise drive a decrease in water vapour transport and limit ice-sheet growth. Thus, the obliquity minimum within isotope stage G6 at 2.7 Myr may have succeeded in beginning the age of intense Northern Hemisphere glaciations specifically because it triggered the development of the subarctic Pacific halocline, which then continued to provide water vapour to boreal North America even as the globally averaged atmosphere became colder and drier.

North Pacific seasonality and glaciation

To test the links among subarctic Pacific stratification, SST and ice-sheet growth, we carry out a suite of experiments with CLIMBER-2, an Earth system model of intermediate complexity^{31,32}. To control stratification, we vary the freshwater input into the subarctic North Pacific region. If this input is reduced from modern forcing by 0.2 Sv ($\sim 20\%$ of the total precipitation over the North Pacific), the model comes to equilibrium with a 'destratified' subarctic Pacific that lacks its modern halocline. An increase in freshwater input is unlikely to have been the specific cause of subarctic Pacific stratification at 2.7 Myr (ref. 13); it merely represents a simple model strategy for changing polar stratification³¹ (see also Supplementary Information).

The 'destratified' and 'stratified' equilibria differ in ways that are consistent with their representation of pre- and post-2.7 Myr conditions, respectively. Relative to the 'destratified' equilibrium, the modern equilibrium maintains much colder winter and spring SSTs in the subarctic Pacific and has significant seasonal sea-ice cover, which the 'destratified' equilibrium lacks. Despite the overall cooling associated with the modern equilibrium, late-summer and autumn SST is actually warmer in this modern equilibrium, which we explain above as the result of reduced thermal inertia associated

with stratification. Relative to the destratified state, the cold spring in the stratified state reduces spring and summer snowmelt (Fig. 2a, b). At the same time, the warm autumn SST of the stratified state maintains the moisture supply to North America (Fig. 2c, d) despite annually averaged cooling.

To assess the significance of the differences between the stratified and destratified states for the build-up of ice sheets in the Northern Hemisphere, we performed additional experiments using an extreme orbital configuration called 'cold orbit'. A high-resolution snow pack model coupled to CLIMBER-2 was used to diagnose the area of permanent snow cover, which can be considered as a minimum footprint for the ice sheets. For the modern climate state (stratified North Pacific) and the 'cold orbit', a large area of North America is perennially covered by snow (Fig. 3b). In contrast, with a destratified North Pacific, the area of perennial snow cover is restricted to the Arctic archipelago and small mountainous areas (Fig. 3a). Growth of the ice sheet provides an additional positive feedback, which explains part of the large temperature difference between the stratified and destratified North Pacific climate states (Fig. 3b, see also Supplementary Information).

A time-evolving experiment simulates the development of stratification at 2.7 Myr (Fig. 1i). The experiment begins at 3.1 million years ago (Fig. 1i) from the destratified state and gradually increases the freshwater flux to the subarctic Pacific at a constant rate of 0.2 Sv per million years. The simulation also includes orbital parameter variation¹. At 2.75 Myr, stratification sets in, winter/springtime (March/April) subarctic Pacific SST drops by $\sim 5^\circ\text{C}$, and summer-time/autumn (August/September) SST increases by $\sim 3^\circ\text{C}$. The timing of the gradual freshwater increase has been optimized to yield stratification at 2.7 Myr. However, the abrupt development of stratification from the gradual change in freshwater input was a natural response of the model, indicating that a 2.7-Myr stratification event may have been a threshold response to a gradual change in forcing.

Previous work on the connection between water vapour supply and Northern Hemisphere glaciation has focused on the North Atlantic. The connection between deep convection and meridional heat transport in the North Atlantic has represented a central motivation for this focus³³. It is obvious that the North Pacific, a large oceanic region that is upstream of North America in

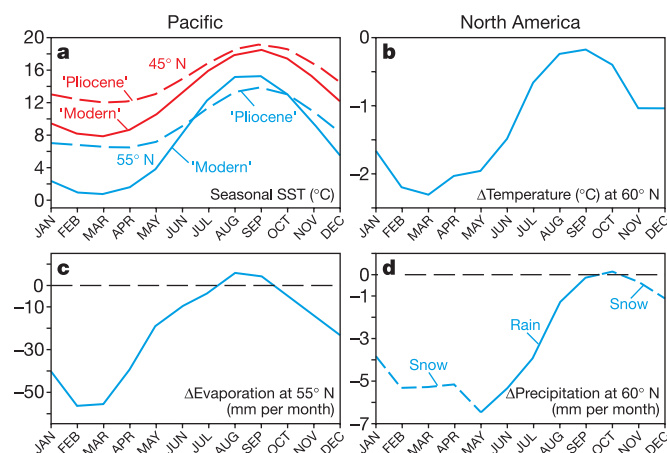


Figure 2 Output for two equilibrium states of the CLIMBER-2 Earth system model. **a**, Seasonal variation in North Pacific SST at 55°N (blue) and 45°N (red). The solid lines correspond to the 'modern' (stratified) equilibrium, the dashed lines to the 'Pliocene' (unstratified) equilibrium state. **b–d**, 'Modern' minus 'Pliocene' differences in seasonal variation of temperature (**b**) and precipitation (as snow and rain) zonally averaged over Northern America at 60°N (**d**), and evaporation from the North Pacific at 55°N (**c**).

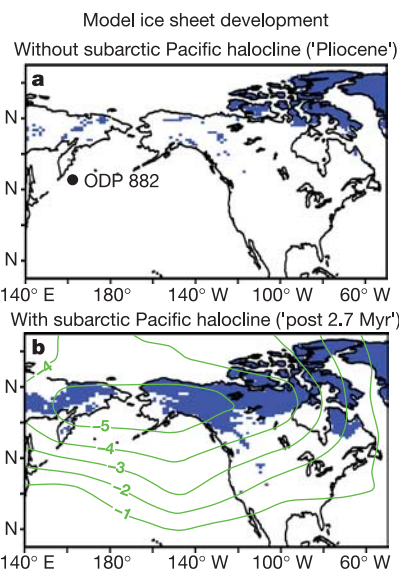


Figure 3 Simulated area of permanent snow cover (shaded) for the 'cold orbit' configuration in the destratified (**a**) and stratified (**b**) equilibria. **b**, Isolines (green) show differences in annual surface air temperature between climate states corresponding to stratified and destratified equilibria under the 'cold orbit' configuration.

atmospheric circulation, could play a critical role in the development of Northern Hemisphere glaciation. Ironically, it was the isolation of the subarctic Pacific surface from the ocean interior that set the stage for major Northern Hemisphere glaciation at 2.7 Myr. □

Received 18 October; accepted 30 December 2004; doi:10.1038/nature03332.

1. Berger, A. & Loutre, M. F. Insolation values for the climate of the last 10 million years. *Quat. Sci. Rev.* **10**, 297–317 (1991).
2. Shackleton, N. J., Hall, M. & Pate, D. Pliocene stable isotope stratigraphy of Site 846. *Proc. ODP Sci. Res.* **138**, 337–357 (1995).
3. Cane, M. A. & Molnar, P. Closing the Indonesian seaway as a precursor to East African aridification around 3–4 million years ago. *Nature* **411**, 157–162 (2001).
4. Ravelo, A. C., Andreasen, D. H., Lyle, M., Olivarez Lyle, A. & Wara, M. W. Regional climate shifts caused by gradual cooling in the Pliocene epoch. *Nature* **429**, 263–267 (2004).
5. Haug, G. H. & Tiedemann, R. Effect of the formation of the Isthmus of Panama on Atlantic Ocean thermohaline circulation. *Nature* **393**, 673–676 (1998).
6. Keigwin, L. D. Isotope paleoceanography of the Caribbean and east Pacific: Role of Panama uplift in late Neogene time. *Science* **217**, 350–353 (1982).
7. Gargett, A. E. Physical processes and the maintenance of nutrient-rich euphotic zones. *Limnol. Oceanogr.* **36**, 1527–1545 (1991).
8. Levitus, S. & Boyer, T. P. *World Ocean Atlas 1994 Vol. 4 Temperature* 129 (NOAA Atlas NESDIS, National Oceanographic Data Center, Silver Spring, USA, 1994).
9. Wong, C. S., Whitney, F. A., Tsoy, I. & Bychkov, A. in *Global Fluxes of Carbon and its Related Substances in the Coastal Sea–Atmosphere System* (eds Tsunogai, S. et al.) 339–344 (Proc. 1994 Sapporo IGBP Symp., M&J International, Yokohama, Japan, 1995).
10. Takahashi, K. Seasonal fluxes of pelagic diatoms in the subarctic Pacific, 1982–1983. *Deep-Sea Res.* **33**, 1225–1251 (1986).
11. Ohkouchi, N., Kawamura, K., Kawahata & Okada, H. Depth ranges of alkenone production in the central Pacific Ocean. *Glob. Biogeochem. Cycles* **13**, 695–704 (1999).
12. Pagani, M., Freeman, K. H., Ohkouchi, N. & Caldeira, K. Comparison of water column $[\text{CO}_2\text{aq}]$ with sedimentary alkenone-based estimates: A test of the alkenone- CO_2 proxy. *Paleoceanography* **17**(4), doi:10.1029/2002PA000756 (2002).
13. Sigman, D. M., Jaccard, S. & Haug, G. H. Polar ocean stratification in a cold climate. *Nature* **428**, 59–63 (2004).
14. Haug, G. H., Sigman, D. M., Tiedemann, R., Pedersen, T. F. & Sarnthein, M. Onset of permanent stratification in the subarctic Pacific Ocean. *Nature* **40**, 779–782 (1999).
15. Maslin, M. A. et al. Northwest Pacific Site 882: The initiation of Northern Hemisphere glaciation. *Proc. ODP Sci. Res.* **145**, 315–333 (1995).
16. Shackleton, N. J. et al. Oxygen isotope calibration of the onset of ice-rafting and history of glaciation in the North Atlantic region. *Nature* **307**, 620–623 (1984).
17. Brassell, S. C., Eglinton, G., Marlowe, I. T., Pflaumann, U. & Sarnthein, M. Molecular stratigraphy: A new tool for climatic assessment. *Nature* **320**, 129–133 (1986).
18. Prahl, F. G. & Wakeham, S. G. Calibration of unsaturation patterns in long-chain ketone compositions for paleotemperature assessment. *Nature* **330**, 367–369 (1987).
19. Müller, P. J., Kirst, G., Ruhland, G., von Storch, I. & Rosell-Mele, A. Calibration of the alkenone paleotemperature index U_{37}^K based on core-tops from the eastern South Atlantic and the global ocean (60°N–60°S). *Geochim. Cosmochim. Acta* **62**, 1757–1772 (1998).
20. Sachs, J. P. et al. Alkenones as paleoceanographic proxies. *Geochem. Geophys. Geosyst.* **1**, 1–13 (2000).
21. Volkman, J. K. Ecological and environmental factors affecting alkenone distributions in seawater and sediments. *Geochem. Geophys. Geosyst.* **1**, 1–12 (2000).
22. Bard, E. Comparison of alkenone estimates with other paleotemperature proxies. *Geochem. Geophys. Geosyst.* **2**, 1–12 (2001).
23. Rosell-Mele, A., Carter, J., Parry, A. & Eglinton, G. Novel procedure for the determination of the U_{37}^K in sediment samples. *Anal. Chem.* **67**, 1283–1289 (1995).
24. Grimalt, J. O. et al. Modification of the C_{37} alkenone and alkenoate composition in the water column and sediments: Possible implications for sea surface temperature estimates in paleoceanography. *Geochem. Geophys. Geosyst.* **1**, 1–20 (2000).
25. Prahl, F. G., Wolfe, G. V. & Sparrow, M. A. Physiological impacts on alkenone paleothermometry. *Paleoceanography* **18**(1052), doi:10.1029/2002PA000853 (2003).
26. Ohkouchi, N., Eglinton, T. I., Keigwin, L. D. & Hayes, J. M. Spatial and temporal offsets between proxy records in a sediment drift. *Science* **298**, 1224–1227 (2002).
27. Juillet-Leclerc, A. & Labeyrie, L. Temperature dependence of the oxygen isotope fractionation between diatom silica and water. *Earth Planet. Sci. Lett.* **84**, 69–74 (1987).
28. Shemesh, A., Charles, C. D. & Fairbanks, R. G. Oxygen isotopes in biogenic silica: global changes in ocean temperature and isotopic composition. *Science* **256**, 1434–1436 (1992).
29. Brandriss, M. E., O'Neil, J. R., Edlund, M. B. & Stoermer, E. F. Oxygen isotope fractionation between diatomaceous silica and water. *Geochim. Cosmochim. Acta* **62**, 1119–1125 (1998).
30. Koster, R. et al. Global sources of local precipitation as determined by the NASA/GISS GCM. *Geophys. Res. Lett.* **13**, 121–124 (1986).
31. Ganopolski, A., Rahmstorf, S., Petoukhov, V. & Claussen, M. Simulation of modern and glacial climates with a coupled global model of intermediate complexity. *Nature* **391**, 351–356 (1998).
32. Petoukhov, V. et al. CLIMBER-2: A climate system model of intermediate complexity. Part I: Model description and performance for present climate. *Clim. Dyn.* **16**, 1–17 (2000).
33. Broecker, W. S. Thermohaline circulation, the Achilles heel of our climate system: will man-made CO_2 upset the current balance? *Science* **278**, 1582–1588 (1997).
34. Tiedemann, R., Sarnthein, M. & Shackleton, N. J. Astronomic timescale for the Pliocene Atlantic $\delta^{18}\text{O}$ and dust flux records of ODP Site 659. *Paleoceanography* **9**, 619–638 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Sarnthein, H. Thierstein, M. Zhao and S. Honjo for discussions. J. Barron, J. Onodera and K. Takahashi provided insight on diatoms *C. marginatus* and *C. radiatus* and their seasonal fluxes in the North Pacific, and H. Sloane helped with the diatom $\delta^{18}\text{O}$ analyses. We thank the Ocean Drilling Program (ODP) and the scientific party and crew of ODP Leg 145 for their efforts in the drilling of Site 882. This work was supported by the Deutsche Forschungsgemeinschaft (DFG), Schweizer Nationalfonds (SNF), the US National Science Foundation (NSF) and BP and the Ford Motor Company through the Princeton University Carbon Mitigation Initiative.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.H.H. (haug@gfz-potsdam.de).

The zinc finger transcription factor *Th-POK* regulates CD4 versus CD8 T-cell lineage commitment

Xiao He^{1*}, Xi He^{1*}, Vibhuti P. Dave^{1†}, Yi Zhang¹, Xiang Hua¹, Emmanuelle Nicolas¹, Weihong Xu², Bruce A. Roe² & Dietmar J. Kappes¹

¹Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, Pennsylvania 19111, USA

²Department of Chemistry And Biochemistry, University of Oklahoma, 620 Parrington Oval, Room 208, Norman, Oklahoma 73019, USA

* These authors contributed equally to this work

† Present address: Lymphocyte Development Laboratory, Room 1045, I.R.C.M., 110 Pine Avenue West, Montreal, Quebec H2W 1R7, Canada

Development of immature T-cell precursors (thymocytes) to either the CD4 helper or CD8 killer T-cell lineages correlates precisely with their T-cell receptor specificity for major histocompatibility complex class II or class I molecules, respectively, indicating that the process is carefully regulated. Although intensively studied owing to its importance in determining the composition of the mature T-cell compartment and as a general model of binary lineage decisions, the underlying molecular pathways remain obscure. We have previously reported a spontaneous mouse mutant (HD (helper deficient) mice) in which lineage commitment is specifically perturbed without affecting positive selection. Here we show that a point mutation in the zinc finger transcription factor *Th-POK* (T-helper-inducing POZ/Krüppel-like factor) is responsible for redirection of class-II-restricted thymocytes to the CD8 lineage in HD mice. Furthermore, we demonstrate that constitutive expression of this factor during thymic development leads to redirection of class-I-restricted thymocytes to the CD4 lineage, indicating that *Th-POK* is a master regulator of lineage commitment.

Developing $\alpha\beta$ T cells progress through three major stages in the thymus, defined by differential expression of the CD4 and CD8 co-receptor molecules; that is, CD4[−]CD8[−] (double negative), CD4⁺CD8⁺ (double positive) and CD4⁺CD8[−] or CD4[−]CD8⁺ (single positive). The double-positive to single-positive transition depends on productive rearrangement of both α - and β -subunits of the T-cell receptor (TCR) and engagement of the complete $\alpha\beta$ TCR by intrathymic ligands (positive selection). Simultaneously, thymocytes diverge into the functionally distinct T-helper and T-killer lineages, defined by expression of CD4 and CD8, respectively. Mature T cells show an almost perfect correlation between CD4 or CD8 expression and their TCR specificity towards class II or class I major histocompatibility complex (MHC) molecules, respectively. Alternative instructive and stochastic/selective models have been proposed to explain this marked correlation (for recent reviews see refs 1, 2). Current thinking favours a quantitative version of the instructive model, whereby lineage choice is determined by the relative strength or duration of TCR engagement^{3–9}; however, the intracellular pathways that are involved remain unknown.

Progress in the field has been hindered because lineage commitment is so intimately tied to the process of positive selection that it is difficult to study in isolation. Hence, no specific pathways have been identified that are required for lineage commitment but not positive selection. Recently, we identified a spontaneous recessive mutation in mice, the HD mutation, which appeared to identify a genetic locus specifically required for lineage commitment¹⁰. Notably, this mutation caused redirection of all class-II-restricted thymocytes to the CD8 lineage¹¹. The existence of such a mutation demonstrated a mechanistic distinction between the pathways governing lineage commitment and positive selection, and provided a uniquely informative tool for dissecting the intracellular pathways that govern lineage commitment. Here we identify the HD locus as the zinc finger transcription factor *Th-POK*. We show that its expression in the thymus is normally restricted to the CD4 lineage, and that constitutive expression leads to redirection of class-I-restricted thymocytes to the CD4 lineage.

No primary defect in CD4 or CD8 expression in HD mice

One possible basis for lineage redirection in HD^{−/−} mice is a primary defect in expression of one of the co-receptor molecules. For instance, in CD4^{−/−} mice a substantial proportion of class-II-restricted thymocytes undergoes redirected development to the CD8 lineage^{3,12}. In HD^{−/−} mice, class-II-restricted thymocytes undergo aberrant downregulation of CD4 and upregulation of CD8, either of which could be the primary cause of altered lineage development. To test whether loss of CD4 expression is causing the defect, a constitutively expressed CD4 transgene¹³ was introduced onto the HD^{−/−} background. Fluorescence-activated cell sorting (FACS) comparison of thymocytes from HD^{−/−} and HD^{+/−} mice expressing the CD4 transgene shows that CD4 development is not restored (Fig. 1a). HD^{+/−} mice bearing the CD4 transgene generate mature thymocytes and peripheral T cells of both lineages, that is, single-positive CD4⁺ and double-positive cells (the latter corresponding to CD8-committed cells), whereas HD^{−/−} mice generate only double-positive thymocytes (CD8 committed). To test the converse possibility, that aberrant upregulation of CD8 might be the primary cause of the HD defect, HD^{−/−} mice were crossed to CD8a^{−/−} mice. Significantly, HD^{−/−}CD8a^{−/−} double-deficient animals did not exhibit restoration of mature, single-positive CD4⁺ thymocytes or peripheral CD4⁺ T cells, indicating that aberrant CD8 upregulation is not the primary defect in HD^{−/−} mice (Fig. 1a). A substantial population of mature double-negative T cells is observed in the thymus and periphery of these mice, presumably consisting of class-II-restricted cells that continue to develop to the CD8 lineage but cannot express surface CD8 (Fig. 1a). This analysis demonstrates that aberrant expression of CD4 or CD8 cannot by itself be responsible for altered lineage commitment of class-II-restricted thymocytes in HD^{−/−} mice. Instead, CD4 and CD8 expression are regulated coordinately in a manner consistent with normal commitment to the CD8 lineage. In further support of normal CD8 commitment, redirected class-II-restricted thymocytes from HD^{−/−} $\beta 2m^{-/-}$

mice show upregulation of the CD8 lineage-specific gene perforin (Fig. 1b).

No detectable TCR signalling defect in HD mice

Given the important role of the TCR in lineage commitment, it is possible that the HD phenotype reflects a defect in TCR signalling. We have previously reported that several proximal events in the TCR signalling cascade including phosphorylation of CD3- ζ , ZAP-70 and p56^{lck} are normal in HD^{-/-} thymocytes¹¹. We have now found that TCR-mediated induction of Ca²⁺ flux, and phosphorylation of the MAP kinases Erk, Jnk and p38, are also unaltered (data not shown). Positive and negative selection provide sensitive readouts of the efficiency of TCR signalling *in vivo*. If redirection of class-II-restricted thymocytes in HD^{-/-} mice is due to an overall deficiency in TCR signalling, then selection defects should also be apparent for

class-I-restricted TCRs. To test this, the class-I-restricted HY TCR—which recognizes a male-specific antigen and mediates positive and negative selection on female and male backgrounds, respectively—was introduced onto the HD^{-/-} background. Positive selection was examined in RAG^{-/-} females to preclude expression of endogenous TCR specificities. No significant difference was detected between HY⁺ HD^{-/-} and HY⁺ HD^{+/+} female mice in terms of the number of single-positive CD8⁺ thymocytes generated, indicating that positive selection of class-I-restricted TCRs is unaffected by the HD mutation (Supplementary Fig. S1). Similarly, HY-mediated negative selection was normal in HD^{-/-} male mice, as demonstrated by severe reductions of double-positive and single-positive thymic populations (Supplementary Fig. S1). Altogether, these data reinforce the conclusion that the HD mutation only affects lineage commitment. If the defect lies in a TCR-mediated signalling

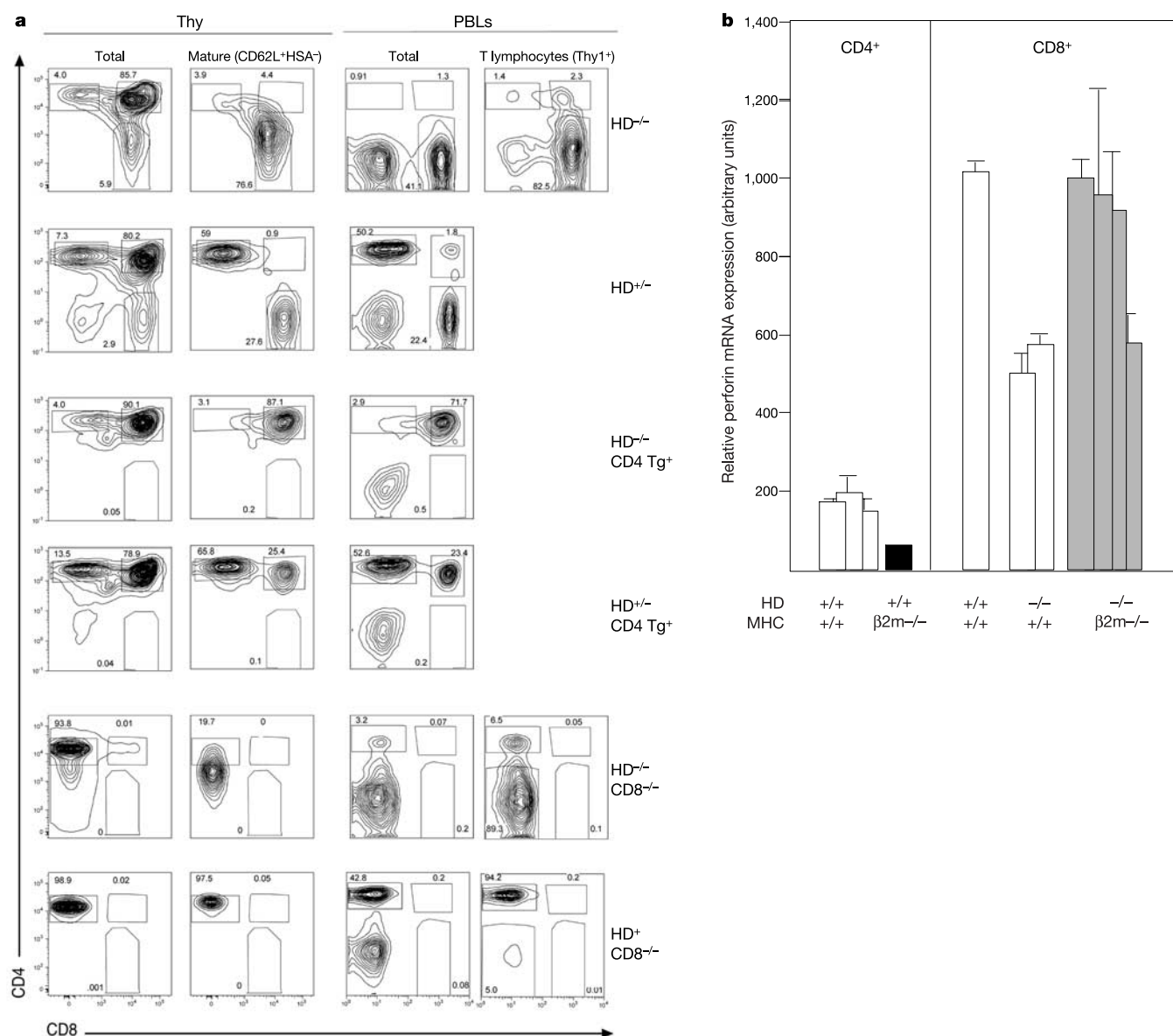


Figure 1 Constitutive CD4 expression or ablation of CD8 does not correct the HD phenotype. **a**, FACS analysis of thymocytes and peripheral blood lymphocytes (PBLs) from HD^{-/-} and HD^{+/+} mice either expressing a constitutive CD4 transgene (Tg⁺) or lacking CD8 expression, stained with anti-CD4, anti-CD8, anti-HSA (heat-stable antigen or CD24) and anti-CD62L. Gating on CD62L⁺ HSA⁻ thymic subsets reveals fully mature thymocytes (second column), and gating on Thy1⁺ PBLs reveals double-negative T cells

in HD^{-/-} CD8^{-/-} mice. At least two mice of each genotype were analysed and all showed normal thymic cellularity ($0.7\text{--}2.3 \times 10^8$). Values of boxed regions are percentages. **b**, Real-time RT-PCR analysis of perforin mRNA levels in sorted single-positive thymocytes from indicated mice. Duplicate bars represent different mice and error bars represent standard deviations.

pathway, it must be a specialized pathway not required for positive or negative selection.

Genetic mapping of the HD defect

Lacking obvious candidate genes, we decided to map the HD locus genetically. For this purpose, HD^{-/-} mice, which arose from a complex mixture of BALB/c, C57BL/6, 129/Sv and C3H strains, were crossed to wild mouse subspecies *Mus musculus castaneus* and *Mus musculus molossinus*, from which they were expected to differ at most polymorphic markers. The resulting HD^{+/-} F₁ animals were backcrossed to HD^{-/-} mice to generate a panel of >2,000 N2 hybrid mice. All N2 animals were typed for the presence of peripheral CD4⁺ T cells by FACS analysis, and a subset of 635 mice (~25%) was identified that exhibited the HD phenotype. Because the HD mutation is recessive, these animals were by definition homozygous for the mutant HD allele (HD^{-/-}). By typing a fraction of these HD^{-/-} N2 animals for a panel of chromosome-specific markers, the HD locus was mapped to mouse chromosome 3 (Fig. 2a). Narrower resolution was obtained by typing all 635 N2 HD^{-/-} animals for a panel of closely spaced chromosome-3-specific markers. Finally, analysis of the 20 most informative recombinant mice narrowed the location of the HD mutation to an interval of <1 cM between markers D3Mit49 and D3Mit341 (Fig. 2b). On the basis of genetic mapping data for the mouse and genomic sequence data for the syntenic human region available at the time, about 30 genes could be tentatively identified as mapping within the HD region. Polymerase chain reaction (PCR) analysis of HD^{-/-} genomic DNA proved positive for all of these genes, ruling out a major genomic deletion.

Identification of the HD locus

To map the mutation more precisely, we used a bacterial artificial chromosome (BAC) transgenic complementation approach using a set of eight BAC clones, which collectively spanned the entire genetically defined HD region, as defined by D3Mit341 and D3Mit49 markers (Fig. 3a). BAC transgenic mice were crossed to HD^{-/-} mice to generate BAC⁺ HD^{-/-} N2 animals. On the basis of the restoration of single-positive CD4 cell numbers in the thymus and periphery, two of seven BACs (A126P10 and A368D24) were able to correct the HD phenotype (Fig. 3a; see also Supplementary Table S1 and Supplementary Fig. S2). Not all lines that carried these

two BACs mediated phenotypic rescue, probably due to fragmentation of large BAC clones during DNA preparation or injection. The ~160-kilobase (kb) region of overlap between these two BACs, which must logically contain the HD locus, included ten known intact genes (Fig. 3a).

To confirm and refine the BAC complementation analysis, a series of truncated BAC derivatives was generated from the rescuing BAC clones and again used for transgenic complementation (Fig. 3b). Eight deletion variants rescued the HD phenotype, and all of them shared a common region of 33 kb ('HD region' in Fig. 3b). This region contained two intact genes: *Lep503*, which encodes a protein of unknown function and is expressed in epithelial cells, and *Th-POK* (also known as *cKrox*, *Zfp67*), a transcription factor (Figs 3b and 4)¹⁴. Clones 368.6 and 368.28, which did not complement the HD phenotype, contained intact *Lep503* but a truncated *Th-POK* gene, resulting in loss of the last 47 amino acids (Figs 3b and 4a), implying that *Th-POK* is the HD gene.

To confirm this, we searched for mutations in *Th-POK* transcripts from HD^{-/-} mice. Complementary DNAs from HD mutant mice showed a specific A to G change at nucleotide position 1165 within the *Th-POK* coding region, resulting in an Arg to Gly substitution at amino acid position 389 (Fig. 4c). This mutation was confirmed in multiple, independently cloned cDNAs as well as in genomic DNA from HD mice. All wild-type *Th-POK* sequences determined by us or found in public databases encode an Arg at this position, including mouse sequences from BALB/c, C57BL/6, FVB and CZECH II strains, as well as homologues from human and five other vertebrate species, implying that this residue is functionally critical. The Arg to Gly substitution occurs within the second of four zinc finger domains of *Th-POK* (Fig. 4b), and affects a residue predicted to interact directly with DNA (according to the model of ref. 15). It has been shown that mutagenesis of key individual residues within zinc finger domains can abolish their DNA binding ability¹⁶, suggesting that the HD defect results from failure of *Th-POK* to bind to a key regulatory site(s).

Th-POK is expressed specifically in the CD4 lineage

On the basis of our previous demonstration that the HD defect is intrinsic to the T-cell lineage¹¹, we predicted that *Th-POK* should be expressed in developing thymocytes. Initial northern blot analysis confirmed that *Th-POK* RNA is expressed in the thymus as well as

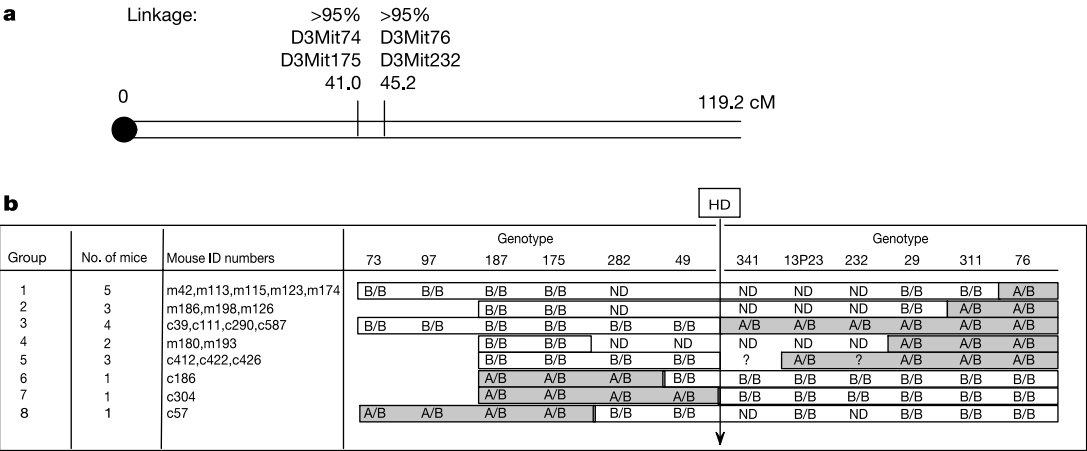


Figure 2 Genetic mapping of the HD defect. **a**, Genetic linkage analysis demonstrating a high degree of co-segregation of the HD allele with markers on mouse chromosome 3. **b**, Fine mapping of the HD locus using closely spaced chromosome 3 markers on a subset of highly informative recombinant mice. Heterozygous or homozygous typing outcomes are designated by A/B and B/B, respectively (the A/A haplotype is impossible in this

breeding scheme). No data (ND) indicates instances when polymorphic markers do not distinguish between BALB/c and either *M. m. molossinus* or *M. m. castaneus* haplotypes. Question marks indicate ambiguous typing results. The c and m prefixes in the third column indicate *M. m. castaneus* and *M. m. molossinus*, respectively. Numbers in the genotype column refer to the relevant MIT marker (for example, 73 indicates D3Mit73).

many other tissues (Supplementary Fig. S3), consistent with a previous report¹⁴ and data from expressed sequence tag (EST) databases. Real-time PCR with reverse transcription (RT-PCR) analysis of sorted wild-type thymic subsets revealed a marked stage- and lineage-specific expression pattern, such that *Th-POK* transcripts were restricted to mature single-positive CD4⁺ and intermediate CD4⁺CD8^{lo} subsets (Fig. 5). The latter subset contains precursors of both the CD4 and CD8 lineages. To distinguish between class-I- and class-II-restricted cells at this stage, we used mice in which only T cells of one or the other specificity can be selected; that is, mice deficient for either class I or class II MHC or expressing a particular TCR transgene. Notably, *Th-POK* messenger RNA expression was preferentially upregulated in class-II-restricted CD4⁺CD8^{lo} cells, consistent with a critical early role in lineage specification (Fig. 5). Interestingly, in HD^{-/-}β2m^{-/-} mice, in which all selected cells are class-II-restricted, *Th-POK* is upregulated at the CD4⁺CD8^{lo} stage but returns to background levels in mature single-positive CD8 thymocytes. This correlation between loss of CD4 expression and *Th-POK* downregulation suggests that *Th-POK* expression is controlled by TCR-mediated signalling.

Ectopic *Th-POK* drives CD4 development

The lineage-specific expression pattern of *Th-POK* suggested that its induction might be sufficient for CD4 commitment. In this case, enforced constitutive expression of *Th-POK* should direct all thymocytes to the CD4 lineage. To test this, we used two complementary approaches: retroviral transduction of bone marrow and

stable transgenesis using T-cell specific promoters. The retroviral approach allows very high transgene expression in the haematopoietic compartment and identification of cells expressing the transgene due to an associated green fluorescent protein (GFP) marker. Wild-type and mutant (Arg389Gly) *Th-POK* cDNAs in the MigR1 retroviral vector¹⁷ were transduced into HD^{+/+} or HD^{-/-} bone marrow and then transferred into RAG^{-/-} recipients, which lack endogenous T cells. After reconstitution of the haematopoietic compartment, FACS analysis was carried out on GFP⁺ lymphocytes from thymus and peripheral blood. Notably, overexpression of wild-type *Th-POK* mediated exclusive generation of single-positive CD4⁺ thymocytes and peripheral T cells in both HD^{-/-} and wild-type backgrounds (Fig. 6a). Single-positive CD4⁺ thymocytes exhibited a normal proportion of mature TCR^{hi}CD69^{lo} cells. The fact that almost all single-positive CD4⁺ thymocytes were TCR^{hi} indicates that they underwent normal positive selection, which requires TCR expression and engagement, and implies that *Th-POK* overexpression cannot mediate development to the single-positive CD4⁺ stage without positive selection. In contrast, mutant *Th-POK* had no effect on thymic development in either wild-type or HD^{-/-} backgrounds. The ability of the wild-type but not mutant forms of *Th-POK* to correct the HD defect demonstrates that the Arg389Gly mutation, rather than another uncharacterized mutation linked to the *Th-POK* locus, is responsible for the HD defect. The inability of mutant *Th-POK* to shift development of wild-type thymocytes towards the CD8 lineage, even when expressed at very high levels under the control of retroviral

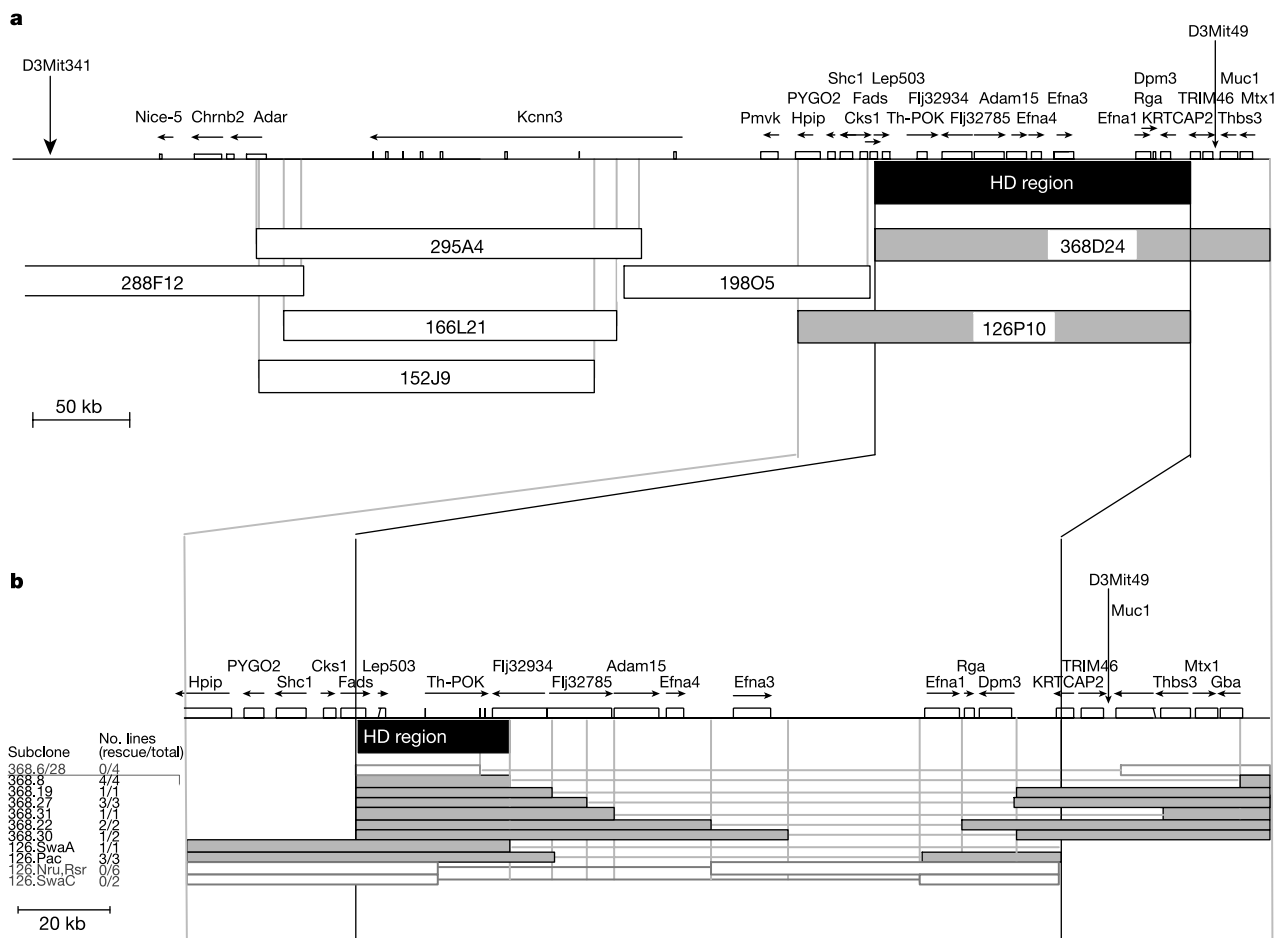


Figure 3 BAC complementation of the HD phenotype. **a**, Map of genetically defined HD region with full-length BAC clones superimposed. Grey shading indicates clones that show rescue. **b**, Rescue by truncated BAC subclones (internal deletions indicated by thin

lines). Grey shading indicates clones that show rescue. The minimum HD region as defined by full-length BAC and BAC subclone rescue is indicated in each panel as 'HD region'.

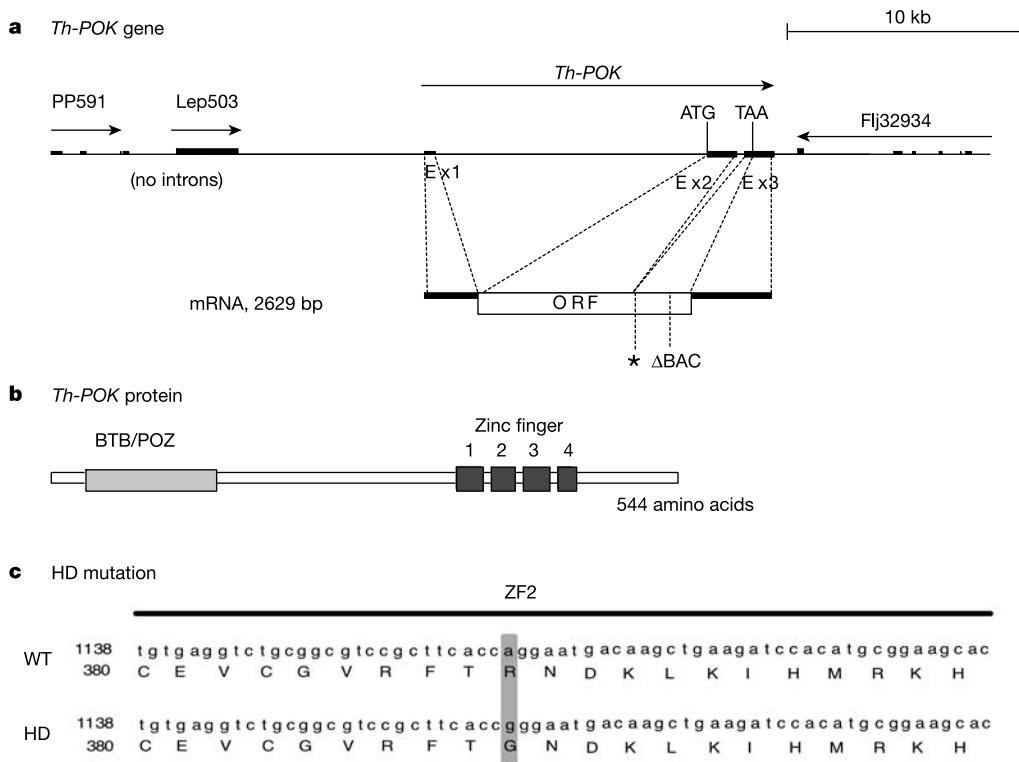


Figure 4 The HD mutation is located in the second zinc finger domain of *Th-POK*. **a, b**, Diagram of *Th-POK* gene organization (**a**) and protein domain structure (**b**). Location of the HD mutation is indicated by an asterisk. Breakpoint of the non-rescuing BAC

subclone 368.6/28 is indicated as 'ΔBAC'. **c**, Nucleotide and amino acid sequence of the second zinc finger domain showing the location of the HD mutation.

enhancers, argues strongly that the mutant form is functionally inert and unable to compete with the wild-type form.

The above experiments show that constitutive *Th-POK* expression prevents development of single-positive CD8⁺ thymocytes; however, it is not clear whether this effect is mediated during

thymic development or earlier in haematopoiesis. In addition, the proportion of GFP⁺ thymocytes and the levels of GFP they express are substantially lower for wild-type compared with mutant *Th-POK* constructs (data not shown), suggesting that premature or excessive expression of wild-type *Th-POK* could partly block T-cell

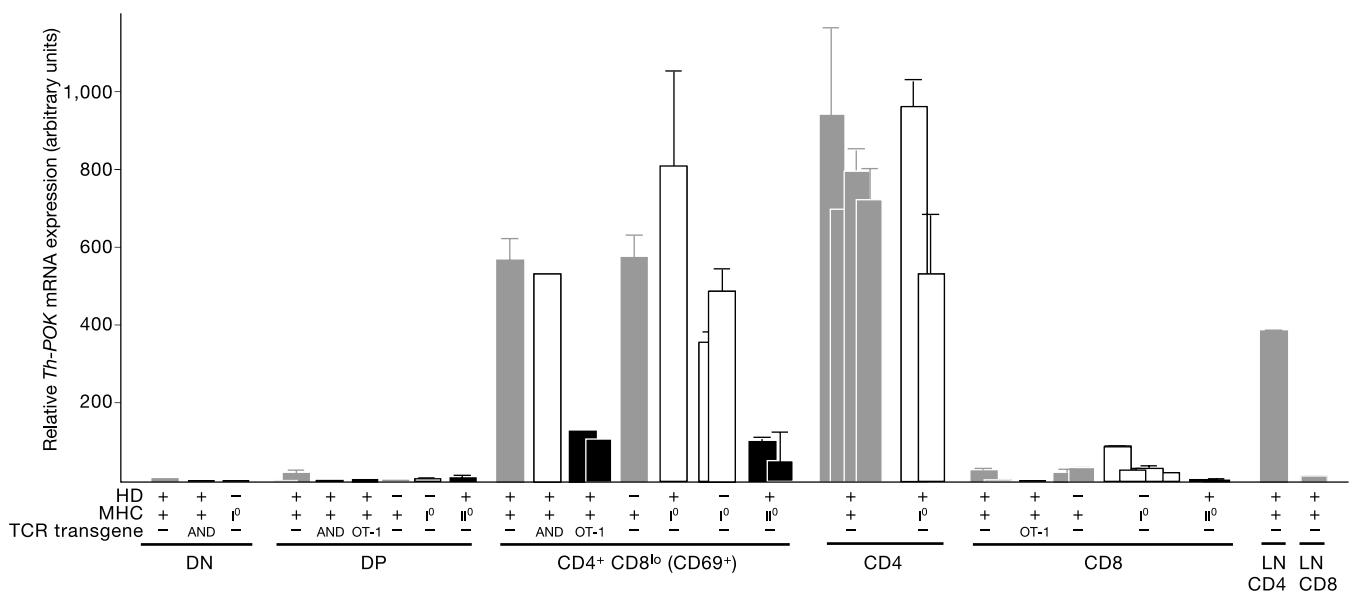


Figure 5 Lineage- and stage-specific expression of *Th-POK* during thymic development. *Th-POK* expression as measured by real-time RT-PCR analysis is shown for sorted thymocyte subsets from indicated mice. Samples from mice in which only class-I- or class-II-restricted thymocytes can be positively selected are indicated by black or white bars, respectively, whereas those from mice with a normal TCR repertoire are indicated in

grey. Duplicate bars for sorted subsets of the same genotype represent different mice. I⁰, MHC class-I-deficient; II⁰, MHC class-II-deficient; DN, double negative; DP, double positive. AND and OT-1 are MHC class-II- and class-I-restricted TCR transgenes, respectively. Error bars represent standard deviations.

development. To restrict *Th-POK* transgene expression to the T-cell lineage and to more physiological levels, we generated stable transgenic lines in which wild-type *Th-POK* was controlled by CD2 regulatory elements ((WT)Th-POK^{CD2})¹⁸. This transgene closely reproduced the key results obtained from retroviral transduction; that is, it restored development of CD4⁺ T cells in an HD^{-/-} background and abolished single-positive CD8⁺ cells in the thymus and periphery of both HD^{-/-} and HD^{+/-} backgrounds (Fig. 6b). However, the (WT)Th-POK^{CD2} transgene did not cause any change in thymic cellularity, indicating that the level of *Th-POK*

expression supported by the CD2 vector does not impair T-cell development. Single-positive CD4⁺ thymocytes were generated in normal numbers, and contained a normal proportion of mature TCR^{hi} CD69^{lo} cells. Similar results were obtained for several independent transgenic lines, in which *Th-POK* expression was controlled either by CD2 or CD4 regulatory elements (data not shown).

To determine whether the absence of single-positive CD8⁺ cells in these mice reflected redirection of class-I-restricted thymocytes to the CD4 lineage, the (WT)Th-POK^{CD2} transgene was crossed to a class-I-restricted TCR transgene, OT-1. Notably, single-positive

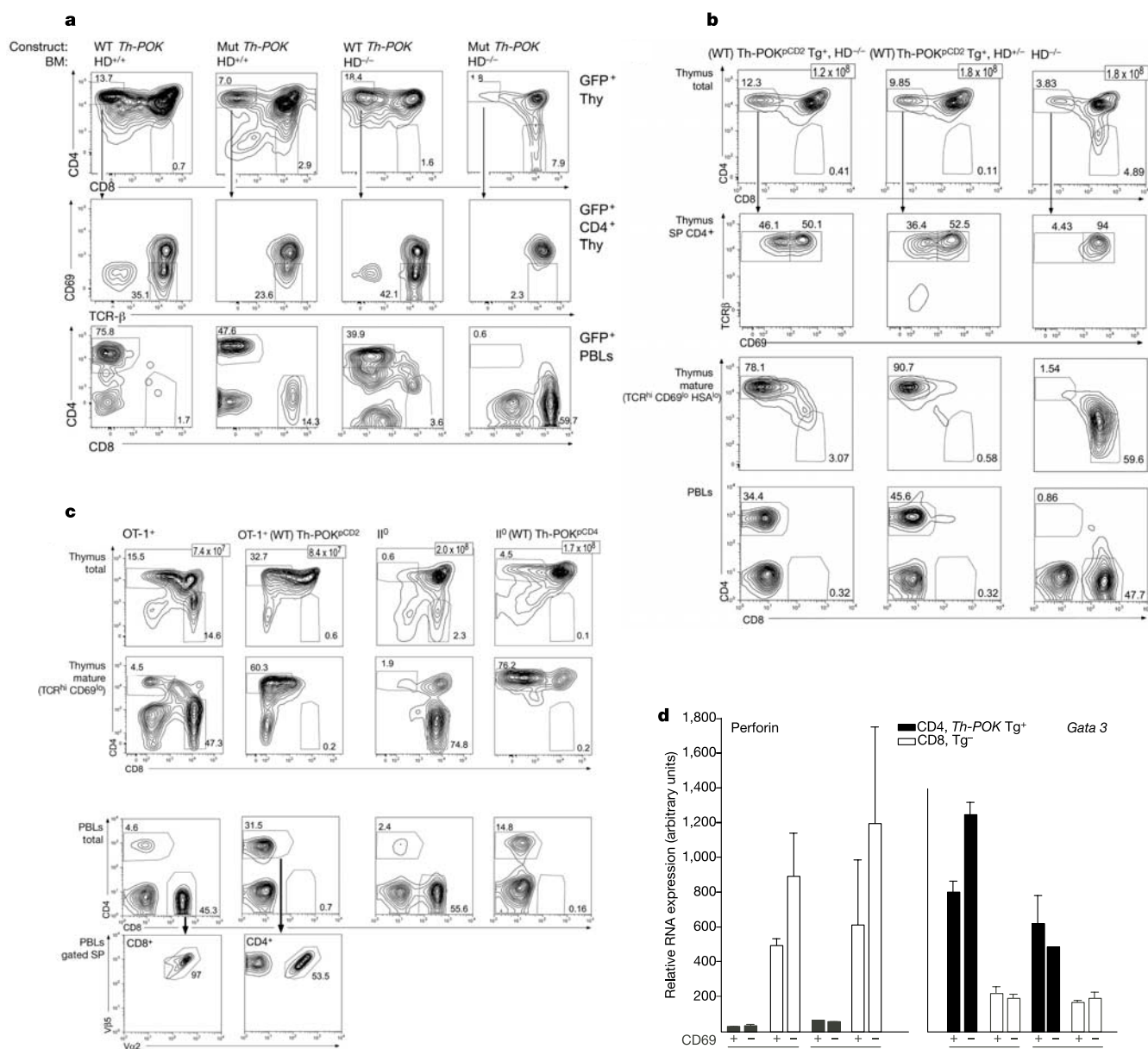


Figure 6 Ectopic expression of *Th-POK* causes redirection of class-I-restricted thymocytes to the CD4 lineage. **a**, FACS analysis of GFP⁺ thymocytes and PBLs from RAG^{-/-} mice, 5 weeks after reconstitution with wild-type or HD^{-/-} bone marrow transduced with either wild-type or mutant *Th-POK* retroviral constructs, as indicated. **b**, FACS analysis of thymocytes and PBLs from HD^{-/-} or HD^{+/-} mice expressing a wild-type *Th-POK* transgene. SP, single positive. **c**, FACS analysis of thymocytes and PBLs from OT-1 transgenic or MHC class-II-deficient (II⁰) mice in the presence or absence of a *Th-POK* transgene. Thymocytes were stained with anti-CD4, anti-CD8, anti-TCR-β and anti-CD69, and PBLs with anti-CD4 and anti-CD8 antibodies. TCR and CD69 expression

is shown for gated CD4⁺ CD8⁻ thymocytes in the second rows of **a** and **b**. CD4 and CD8 expression is shown for gated, mature TCR^{hi} CD69⁻ thymocytes in the third row of **b** and second row of **c**. OT-1 PBLs were additionally stained with anti-Vα2 and anti-Vβ5 antibodies in **c**. At least two animals of each genotype were analysed with similar results. The number at the upper right of each plot indicates total thymic cellularity. **d**, Real-time RT-PCR analysis of perforin and *Gata3* expression in sorted, single-positive CD4⁺ CD69⁺ and CD69⁻ thymocyte subsets from *Th-POK* transgenic OT-1 or MHC class-II-deficient mice (black bars), or equivalent control mice (white bars). Error bars represent standard deviations.

CD8⁺ thymocytes and peripheral T cells were entirely replaced by single-positive CD4⁺ cells, consistent with redirection (Fig. 6c). In the periphery, most single-positive CD4⁺ cells expressed both transgenic V α 2 and V β 5 chains, indicating that they had been selected on the basis of the OT-1 TCR (Fig. 6c). To obtain independent confirmation of this result, a different wild-type *Th-POK* transgenic line ((WT)Th-POK^{pCD4}) was introduced onto a MHC class-II-deficient background, thereby again limiting the selected TCR repertoire to class-I-restricted specificities. In these mice, only mature single-positive CD4⁺ cells were generated in the thymus and periphery in contrast to non-transgenic MHC class-II-deficient control mice, which produced only single-positive CD8 cells (Fig. 6c). The absolute number of peripheral CD4⁺ T cells in OT-1⁺ and MHC class-II-deficient mice expressing *Th-POK* transgenes was lower than that of CD8⁺ T cells in non-transgenic control mice, indicating a reduced efficiency in thymic emigration and/or peripheral survival of class-I-restricted cells lacking CD8. A selective disadvantage for OT-1⁺ cells lacking CD8 would be expected based on studies in OT-1⁺ CD8^{-/-} mice¹⁹. Finally, to demonstrate by criteria other than co-receptor expression pattern that redirected class-I-restricted thymocytes in *Th-POK* transgenic mice are fully committed to the CD4 lineage, we determined RNA levels for perforin and *Gata3*, which represent markers of the CD8 and CD4 lineages, respectively. Notably, single-positive CD4⁺ cells from *Th-POK* transgenic mice showed absence of perforin expression and upregulation of *Gata3*, as expected for CD4-committed cells (Fig. 6d). The 3–5-fold elevation in *Gata3* expression observed in *Th-POK* transgenic CD4⁺ cells relative to non-transgenic CD8⁺ cells from control mice closely matches the fourfold expression difference previously reported between normal single-positive CD4⁺ and CD8⁺ thymocytes²⁰.

Discussion

Our results show that the presence of functional *Th-POK* in thymocytes undergoing positive selection is necessary and sufficient for commitment to the CD4 lineage, whereas its absence is necessary and sufficient for commitment to the CD8 lineage. The fact that *Th-POK* expression is confined to cells expressing a class-II-restricted TCR specificity as well as CD4 suggests that it is induced by a strong TCR signal, consistent with currently favoured quantitative instructive models of lineage commitment. Full maturation of redirected thymocytes in HD and *Th-POK* transgenic mice also supports an instructive model, albeit indirectly, because such maturation is incompatible with the alternative stochastic/selective model. The fact that *Th-POK* is not involved in other TCR-mediated processes (positive and negative selection) would suggest that it lies on a specialized branch of the TCR signalling pathway.

On the basis of the cellular distribution and structure of the HD product, we have proposed the new designation 'T-helper-inducing POZ/Krüppel-like factor' (*Th-POK*) for this locus (previously called *cKrox*, *cKrox- α* or *Zfp67*). *Th-POK* was first described as an *in vitro* interacting factor and transcriptional repressor of collagen genes (thus 'collagen' *Krox* or *cKrox*)^{14,21}. It belongs to a group of factors related to the *Drosophila* segmentation protein Krüppel, and including *Sp1*, *Egr-1*, *WT-1* and *LKLF*. Within this category, *Th-POK* belongs to a subset containing an amino-terminal BTP/POZ domain, the so-called POK (POZ and Krüppel) proteins, also including *Bcl-6* and *PLZF*. The BTB/POZ domain is responsible for homo- and heterodimerization, transcriptional repression activity and interactions with components of the histone deacetylase complex, thereby providing a functional link to enzymatic activities that regulate chromatin conformation. *Th-POK* is highly conserved between mice and humans (100% identity for the zinc finger region). There are precedents for the critical role of zinc finger proteins in regulating haematopoietic lineage branch points, specifically *Gata1* in the case of erythrocyte development^{22,23} and *Gata3* in the Th1 versus Th2 branchpoint²⁴. *Th-POK*, like *Gata3*, is

widely expressed, and on this basis might be expected to mediate important functions in other tissues. We are currently generating a *Th-POK* knockout mouse to test whether it encodes any additional non-redundant functions.

Recently, two nuclear factors have been reported to be important effectors of CD4 lineage development. *Gata3* is preferentially expressed in the CD4 lineage^{20,25} and is required for CD4 development²⁶, whereas *Runx3* is critical for CD4 silencing²⁷. However, inactivation of either of these genes fails to cause redirection of class-II-restricted cells to the CD8 lineage^{26,28}, suggesting that they do not regulate the commitment process itself. Furthermore, transgenic overexpression of the nuclear high-mobility group (HMG) box protein TOX mediates development of thymocytes exclusively to the CD8 lineage²⁹. However, this occurs only in the absence of TCR engagement, not when thymocytes receive normal TCR signals, arguing against a physiological role in lineage commitment^{29,30}.

The identification of *Th-POK* as a central regulator of lineage commitment provides an important new starting point for analysing the underlying molecular pathways. Future research into the transcriptional control of *Th-POK* itself and the nature of its target genes should contribute to unravelling these pathways further. □

Methods

Mice

CD4 (ref. 13) (provided by S. Reiner) and HY TCR³¹ transgenic lines, as well as HD^{-/-} mice^{10,11}, have been previously described. *M. m. castaneus* and *M. m. molossinus* strains were purchased from Jackson Laboratories. All other transgenic lines were generated by the FCCC Transgenic Facility.

Flow cytometry

Cells were prepared from thymus and peripheral blood, and analysed by flow cytometry according to standard procedures. All antibodies were obtained from BD Pharmingen.

Real-time RT-PCR

Real-time RT-PCR analysis for *Th-POK*, perforin and *Gata3* was carried out according to the probe-based method and analysed by the comparative Ct method (compared to β -actin). Primer and probe sequences are available upon request.

Genetic mapping

HD^{-/-} mice, which arose from mixed C57BL/6, BALB/c, C3H and 129/Sv parentage, were crossed to *M. m. castaneus* or *M. m. molossinus* strains, and the resulting HD^{+/-} N1 animals were again crossed to HD^{-/-} mice, resulting in >2,000 N2 hybrid mice. A total of 635 of these mice were identified as HD^{-/-} and used for PCR-based SSLP linkage analysis (see Supplementary Information).

Genomic sequencing

Sequencing was carried out for three C57BL/6 BAC clones (RPCI-23 295A4, 418C12 and 368D24), which collectively span most of the HD region, confirming the overlap of these three clones, presence of the D3Mit49 marker within 368D24, and presence of many of the genes predicted from BAC end sequencing and genetic data (GenBank accession numbers AC104329, AC104327 and AC104632).

BAC complementation

BAC clones spanning the HD region were identified using the BAC assembly database for the C57BL/6 RPCI-23 BAC library (Genome Sequence Centre, BC Cancer Agency). See Supplementary Information for details.

Transgenic production

BAC clones used for generation of transgenic mice were obtained from the BAC Resource Consortium. Subclones of BACs 126P10 and 368D24 bearing internal deletions were generated by restriction endonuclease treatment followed by re-ligation. The boundaries of all deletions were mapped by PCR analysis and DNA sequencing. For transgenic expression of wild-type *Th-POK* cDNA the full-length cDNA insert was cloned into the hCD2 vector¹⁸ or into a vector containing CD4 promoter/enhancer elements (derived from transgenic construct h in ref. 32).

Retroviral transduction

Complete wild-type and mutant *Th-POK* cDNAs were cloned into the BglII site of the MigR1 vector (gift of W. Pear). Preparation of viral supernatant and transduction of bone marrow was performed essentially as described¹⁷.

Received 23 September; accepted 4 December 2004; doi:10.1038/nature03338.

1. Germain, R. N. T-cell development and the CD4–CD8 lineage decision. *Nature Rev. Immunol.* **2**, 309–322 (2002).
2. Bosselut, R. & Singer, A. CD4/CD8 coreceptors in thymocyte development, selection, and lineage commitment: analysis of the CD4/CD8 lineage decision. *Adv. Immunol.* **83**, 91–131 (2004).

3. Matechak, E. O., Killeen, N., Hedrick, S. M. & Fowlkes, B. J. MHC class II-specific T cells can develop in the CD8 lineage when CD4 is absent. *Immunity* **4**, 337–347 (1996).
4. Itano, A. *et al.* The cytoplasmic domain of CD4 promotes the development of CD4 lineage T cells. *J. Exp. Med.* **183**, 731–741 (1996).
5. Yasutomo, K., Doyle, C., Miele, L., Fuchs, C. & Germain, R. N. The duration of antigen receptor signalling determines CD4+ versus CD8+ T-cell lineage fate. *Nature* **404**, 506–510 (2000).
6. Liu, X. & Bosselut, R. Duration of TCR signaling controls CD4–CD8 lineage differentiation *in vivo*. *Nature Immunol.* **5**, 280–288 (2004).
7. Hernandez-Hoyos, G., Sohn, S. J., Rothenberg, E. V. & Alberola-Ila, J. Lck activity controls CD4/CD8 T cell lineage commitment. *Immunity* **12**, 313–322 (2000).
8. Legname, G. *et al.* Inducible expression of a p56Lck transgene reveals a central role for Lck in the differentiation of CD4 SP thymocytes. *Immunity* **12**, 537–546 (2000).
9. Schmedt, C. *et al.* Csk controls antigen receptor-mediated development and selection of T-lineage cells. *Nature* **394**, 901–904 (1998).
10. Davé, V. P., Allman, D., Keefe, R., Hardy, R. R. & Kappes, D. J. HD mice: a novel mouse mutant with a specific defect in the generation of CD4(+) T cells. *Proc. Natl Acad. Sci. USA* **95**, 8187–8192 (1998).
11. Keefe, R., Dave, V., Allman, D., Wiest, D. & Kappes, D. J. Regulation of lineage commitment distinct from positive selection. *Science* **286**, 1149–1153 (1999).
12. Tyznik, A. J., Sun, J. C. & Bevan, M. J. The CD8 population in CD4-deficient mice is heavily contaminated with MHC class II-restricted T cells. *J. Exp. Med.* **199**, 559–565 (2004).
13. Killeen, N. & Littman, D. R. Helper T-cell development in the absence of CD4–p56lck association. *Nature* **364**, 729–732 (1993).
14. Galera, P., Musso, M., Ducey, P. & Karsenty, G. c-Krox, a transcriptional regulator of type I collagen gene expression, is preferentially expressed in skin. *Proc. Natl Acad. Sci. USA* **91**, 9372–9376 (1994).
15. Klevit, R. E. Recognition of DNA by Cys2,His2 zinc fingers. *Science* **253**, 1367–1393 (1991).
16. Nardelli, J., Gibson, T. J., Vesque, C. & Charnay, P. Base sequence discrimination by zinc-finger DNA-binding domains. *Nature* **349**, 175–178 (1991).
17. Pui, J. C. *et al.* Notch1 expression in early lymphopoiesis influences B versus T lineage determination. *Immunity* **11**, 299–308 (1999).
18. Zhumabekov, T., Corbella, P., Tolaini, M. & Kioussis, D. Improved version of a human CD2 minigene based vector for T cell-specific expression in transgenic mice. *J. Immunol. Methods* **185**, 133–140 (1995).
19. Goldrath, A. W., Hogquist, K. A. & Bevan, M. J. CD8 lineage commitment in the absence of CD8. *Immunity* **6**, 633–642 (1997).
20. Hernández-Hoyos, G., Anderson, M. K., Wang, C., Rothenberg, E. V. & Alberola-Ila, J. GATA-3 expression is controlled by TCR signals and regulates CD4/CD8 differentiation. *Immunity* **19**, 83–94 (2003).
21. Widom, R. L., Lee, J. Y., Joseph, C., Gordon-Froome, I. & Korn, J. H. The hcKrox gene family regulates multiple extracellular matrix genes. *Matrix Biol.* **20**, 451–462 (2001).
22. Tsai, S. F. *et al.* Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature* **339**, 446–451 (1989).
23. Blobel, G. A., Simon, M. C. & Orkin, S. H. Rescue of GATA-1-deficient embryonic stem cells by heterologous GATA-binding proteins. *Mol. Cell. Biol.* **15**, 626–633 (1995).
24. Zheng, W. & Flavell, R. A. The transcription factor GATA-3 is necessary and sufficient for the Th2 cytokine gene expression in CD4 T cells. *Cell* **89**, 587–596 (1997).
25. Hendriks, R. W. *et al.* Expression of the transcription factor GATA-3 is required for the development of the earliest T cell progenitors and correlates with stages of cellular proliferation in the thymus. *Eur. J. Immunol.* **29**, 1912–1918 (1999).
26. Pai, S.-Y. *et al.* Critical roles for transcription factor GATA-3 in thymocyte development. *Immunity* **19**, 863–875 (2003).
27. Taniuchi, I. *et al.* Differential requirements for Runx proteins in CD4 repression and epigenetic silencing during T lymphocyte development. *Cell* **111**, 621–633 (2002).
28. Woolf, E. *et al.* Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. *Proc. Natl Acad. Sci. USA* **100**, 7731–7736 (2003).
29. Aliahmad, P. *et al.* TOX provides a link between calcineurin activation and CD8 lineage commitment. *J. Exp. Med.* **199**, 1089–1099 (2004).
30. Wilkinson, B. *et al.* TOX: an HMG box protein implicated in the regulation of thymocyte selection. *Nature Immunol.* **3**, 272–280 (2002).
31. Kisielow, P., Bluethmann, H., Staerz, U. D., Steinmetz, M. & von Boehmer, H. Tolerance in T cell receptor transgenic mice involves deletion of immature CD4+ CD8+ thymocytes. *Nature* **333**, 742–746 (1988).
32. Sawada, S., Scarborough, J. D., Killeen, N. & Littman, D. R. A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development. *Cell* **77**, 917–929 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge technical assistance with flow cytometry by J. Oesterling, and animal husbandry by members of the Laboratory Animal Facility. V.P.D. was responsible for initial genetic mapping of the HD mutation to a resolution of 4 cM. We thank D. Wiest for comments on the manuscript. This work was supported by National Institutes of Health Grants to D.J.K., an NIH Core Grant, and also an appropriation from the Commonwealth of Pennsylvania.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.J.K. (dj_kappes@fccc.edu).

Structure of an unliganded simian immunodeficiency virus gp120 core

Bing Chen¹, Erik M. Vogan^{1,2}, Haiyun Gong¹, John J. Skehel³, Don C. Wiley^{1,2*} & Stephen C. Harrison^{1,2}

¹Children's Hospital Laboratory of Molecular Medicine, Harvard Medical School, and ²Howard Hughes Medical Institute, 320 Longwood Avenue, Boston, Massachusetts 02115, USA

³National Institute of Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

* Deceased

Envelope glycoproteins of human and simian immunodeficiency virus (HIV and SIV) undergo a series of conformational changes when they interact with receptor (CD4) and co-receptor on the surface of a potential host cell, leading ultimately to fusion of viral and cellular membranes. Structures of fragments of gp120 and gp41 from the envelope protein are known, in conformations corresponding to their post-attachment and postfusion states, respectively. We report the crystal structure, at 4 Å resolution, of a fully glycosylated SIV gp120 core, in a conformation representing its prefusion state, before interaction with CD4. Parts of the protein have a markedly different organization than they do in the CD4-bound state. Comparison of the unliganded and CD4-bound structures leads to a model for events that accompany receptor engagement of an envelope glycoprotein trimer. The two conformations of gp120 also present distinct antigenic surfaces. We identify the binding site for a compound that inhibits viral entry.

The envelope glycoproteins of HIV and SIV are the molecular agents of cell attachment and membrane fusion¹. They are trimeric assemblies of gp160 polypeptide chains, which are cleaved during transport to the surface of an infected cell into two fragments known as gp120 and gp41 (refs 2–4). Cleavage enables the protein to undergo a series of conformational changes when it encounters the receptor for these viruses, CD4, and their co-receptor, CXCR4 or CCR5, on the surface of a suitable host cell^{5–8}. The first of these changes, probably confined largely to gp120, accompanies receptor binding^{9,10}. It stabilizes a conformation with which a co-receptor can then associate^{7,8,11}. Co-receptor binding may induce further changes, leading to dissociation of gp120 from the membrane-anchored gp41. The latter then refolds through a series of steps that lead ultimately to fusion of viral and target-cell membranes^{12,13}.

The structures of gp120 and gp41 at the end of this sequence of conformational changes have been determined in a series of X-ray crystallographic studies^{12–14}. Analysis of the postfusion gp41 structure has been particularly important for deriving a picture of the fusion process and for understanding the mechanism of peptide fusion inhibitors^{13,15,16}. The structure of CD4-bound, HIV-1 gp120, in complex with a monoclonal Fab that recognizes the co-receptor site, has provided a framework for analysing envelope antigenicity^{14,17}. The unliganded, prefusion structure of gp120 and the structure of the prefusion trimer (including the prefusion conformation of gp41) have resisted high-resolution analysis.

The structure of SIV gp120 in an unliganded conformation, described here, now fills one of those lacunae. As in the studies of the CD4-bound structure¹⁴, we have used the gp120 'core', from which two large loops of highly variable sequence, V1–V2 and V3, as well as amino- and carboxy-terminal segments, have been deleted (Fig. 1a). Comparison of the new structure of unliganded gp120 with that of the CD4-bound protein shows that part of the molecule—the 'inner domain' (see below)—undergoes unexpectedly extensive conformational rearrangement upon receptor binding. In the process, coherent sites assemble for CD4 and co-receptor interaction. Because the protein we have crystallized is fully glycosylated, we can visualize directly the extent to which oligosaccharides coat its molecular surface. Knowledge of the unliganded conformation allows us to model how the protein might appear

in a gp120/gp41 trimer and to picture the overall conformational change induced by contact with receptor.

Crystallization of the unliganded SIV gp120 core

Design of the fragment that we crystallized followed closely the strategies that led to crystallization of CD4-bound HIV-1 gp120 (refs 14, 18, 19). The gp120 cores of SIVmac 32H (ref. 20) and HIV-1 HXBc2 have 35% sequence identity and over 70% sequence similarity; alignment of the two sequences is unambiguous. Both have seven disulphide bonds in corresponding positions. Of the 13 glycosylation sites on the SIV protein and 18 on the HIV-1 protein, nine are conserved or shifted by no more than one or two residues. We substituted short linkers (GAG) for the V1–V2 and V3 loops and deleted 43 and 22 residues from the N and C termini, respectively. We obtained crystals as described in the Methods (see also ref. 21). In contrast to reported experience with the HIV-1 gp120 core²², deglycosylation prevented rather than facilitated crystallization. The crystals were small; we could record diffraction only to spacings of about 4 Å. The structure determination (see Methods and ref. 21) relied on two isomorphous derivatives, density averaging among three non-isomorphous crystal forms, and use of SeMet substituted protein to locate methionine residues.

Description of the structure

Overview

High overall sequence conservation between SIV and HIV gp120, their common receptor and co-receptors, and the properties of the SIV gp120 core structure described in this paper all support our interpretation, that most of the differences between our structure and that of CD4-bound HIV-1 gp120 reflect conformational changes induced by receptor binding rather than differences between the HIV-1 and SIV proteins. The unliganded gp120 core has the bipartite character first described for the CD4-bound form (Fig. 1), with inner and outer domains designated in relation to the location of N and C termini. The most extensive structural changes are in the inner domain, for which the sequence is even more conserved than for the outer domain, suggesting that a shared pattern of required conformational changes has constrained sequence variation. Four conserved inner-domain disulphide bonds lock in structural elements that must move with respect to

each other; the more globular outer domain has only three disulphides. Moreover, the conformational differences we observe are all consistent with predictions from studies of CD4 binding to gp120 in solution. For example, residues that make up the chemokine-receptor site, which requires CD4 binding to form^{7,8,23}, are not adjacent in our structure, whereas residues at which mutations generate resistance to a group of entry-inhibiting drugs all face a

well-configured pocket²⁴. Finally, the protein we have crystallized binds tightly to a panel of monoclonal antibodies raised against fusion-competent SIV gp120/gp41 trimer (Supplementary Fig. S1). We therefore refer to the structure reported here as the 'unliganded gp120 core', to distinguish it from the 'CD4-bound gp120 core' previously described¹⁴. In contexts in which we believe it important to remind the reader that one derives from SIVmac 32H and the

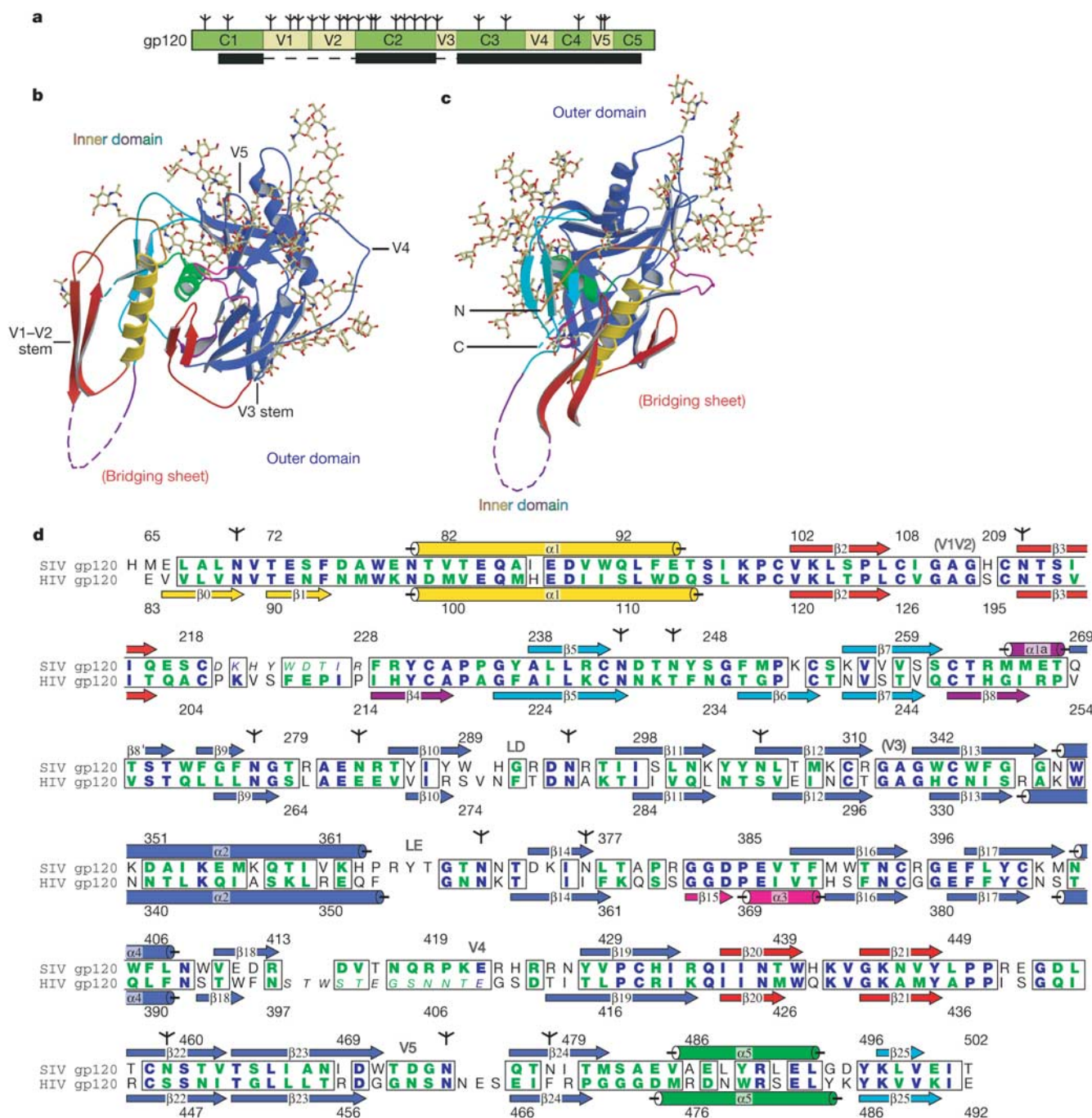


Figure 1 Structure and sequence of SIV gp120 core. **a**, Diagram of sequence elements in gp120 and definition of its core. The branched symbols mark glycosylation sites. **b**, Overall structure of the unliganded SIV gp120 core, with polypeptide chains as ribbon diagram and carbohydrates as stick models. Outer domain is in blue; inner domain, coloured according to substructure (N terminus, orange; α 1, yellow; three-strand sheet, cyan; outer/inner domain transition, purple; α 5, green). The four strands that form the bridging sheet in the CD4-bound conformation are in red and labelled as 'bridging sheet' in parentheses. Disordered residues are shown as dashed lines. The stumps of truncated

variable loops V1, V2 and V3 are indicated and the intact variable loops V4 and V5 are also labelled. **c**, The same as **b**, rotated by 90° about a vertical axis. **d**, Alignment of the gp120 core sequences of SIVmac 32H and HIV-1 HXBc2. Residues in blue are identical; those in green are conservative substitutions; disordered residues are in italic. Glycosylation sites are indicated by tree-like symbols. Secondary structures of both proteins are shown as arrows for β -strands and rods for α -helices. They are coloured following the scheme used for the structure.

other from HIV-1 HXBc2, we add the SIV or HIV-1 designation.

The outer domain has, with some important local exceptions, essentially the same structure in the two states. The inner domain has reorganized markedly. Indeed, the inner domain turns out not to be a single, coherent structure that can shift as a rigid body, but rather a collection of distinct substructures that move independently with respect to each other when the unliganded and CD4-bound structures are compared. These substructures (colour-coded in the figures) include: a four-turn α -helix ($\alpha 1$); a β -ribbon (the 'V1–V2 stem', half of the bridging sheet in the CD4-bound state; $\beta 2$ – $\beta 3$); a 3-strand β -sheet, with two successive strands ($\beta 5$ and $\beta 7$) augmented by the C-terminal strand ($\beta 25$); and a short α -helix ($\alpha 5$) at the outer-domain/inner-domain junction. The N-terminal segment, imperfectly defined at a variable crystal contact, shifts approximately together with the 3-strand sheet. As predicted by Kwong and co-workers¹⁴, the bridging sheet is absent in the unliganded state of gp120. Each of its two β -ribbons is ordered, but a space of 20–25 Å intervenes between them.

The kernel of the outer domain, against which CD4 docks, is an 8-strand, antiparallel β -barrel. The distal end of the barrel extends into a 6-strand β -sheet, which cradles a four-turn α -helix. Some differences between SIV gp120 and HIV-1 gp120 are in the lengths of loops in this domain. Thus, the V4 and V5 loops are slightly shorter in the SIV protein, and loop LE, just C-terminal to the helix and even more variable than V5, is slightly longer (Fig. 2). Two differences between our structure and that of CD4-bound gp120 are clearly owing to CD4 interactions. One is in the connection between a strand of the distal sheet ($\beta 14$) and a strand ($\beta 16$) of the barrel: this 'CD4-binding loop', with a conserved GGDPE sequence, moves when CD4 associates and presents an extended strand ($\beta 15$) for main-chain hydrogen bonding with CD4's C' ridge. The other is in the orientation of the loop ($\beta 20$ – $\beta 21$) that forms one of the two β -ribbons of the bridging sheet (Fig. 2d).

Glycans

Secreted proteins from insect cells generally have mannose-rich (rather than complex) oligosaccharides. We find density for sugars at all 13 glycosylation sites of the SIV gp120 core. Seven of the glycans are clearly $\alpha(1\text{--}6)$ fucosylated, and eight are ordered at least out to the mannose branch site. The longer oligosaccharides form 'glycan clusters' on the molecular surface, probably through hydrogen-bonding networks. A catalogue of the glycan structures is included in ref. 21.

Variable loops

The gp120 core contains neither the V1–V2 nor the V3 loop, but we can infer their approximate locations from the positions of stumps left by their truncation. In the unliganded conformation, their stumps project in almost opposite directions (Figs 1a and 2b), and even allowing for uncertainties about their possible conformations (in SIVmac 32H, the deleted part of V1–V2 contains 99 residues, three disulphide bonds and seven N-linked glycans, which may form a compact structure decorated with sugars, and V3 contains 27 residues and 1 N-linked glycan), it seems unlikely that they make direct contact within a monomer. The V4 loop, disordered in the CD4-bound HIV-1 HXBc2 gp120 core structure¹⁴, lies at a lattice contact in our crystals. It adopts an open conformation that extends away from the body of the outer domain. The amino-acid residues in this loop are almost entirely polar, and we expect that it might be more mobile when not constrained by crystal packing. Unlike the HIV-1 V4 loop, it contains no glycosylation sites. It thus creates an exposed ridge of protein surface above a sea of glycan. Its prominence may explain why neutralizing anti-SIV antibodies often recognize V4 (see refs in Supplementary Fig. S1).

Inner domain: comparison with CD4-bound structure

There are two useful ways to compare the unliganded and CD4-bound structures. If we choose a common orientation for the nearly

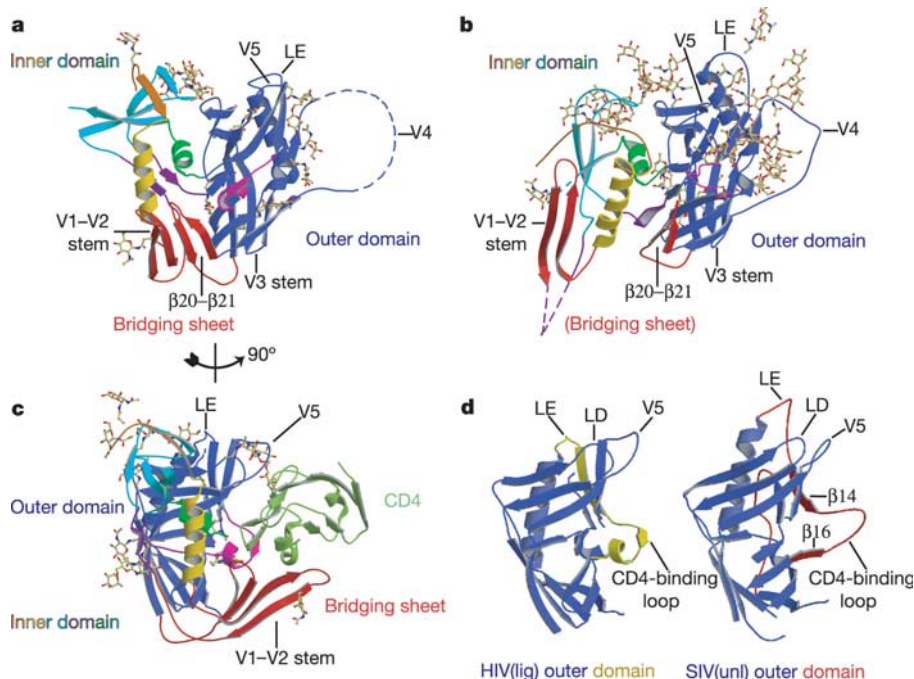


Figure 2 Comparison of glycosylated, unliganded SIV gp120 core with deglycosylated, liganded HIV-1 HXBc2 gp120 core¹⁴. **a**, Conformation of deglycosylated HIV gp120 core, complexed with two-domain CD4 and the Fab fragment of human neutralizing monoclonal 17b, with the outer domain on the right and the inner domain on the left. CD4 and Fab have been omitted for clarity. Twelve disordered V4-loop residues are shown as a dashed line. Colours as in Fig. 1b. **b**, SIV gp120 core structure, in the same view and colours as in

panel **a**. **c**, Interaction of gp120 core and CD4, shown by a side view (90° from the view in panel **a**) of HIV gp120 core complexed with CD4 (light green). **d**, Comparison of outer domains of liganded (lig) HIV gp120 core (left) and unliganded (unl) SIV gp120 core (right). The differences are highlighted in yellow and red, and the CD4-binding loops are labelled. LD and LE are loop designations, following the convention of ref. 14.

invariant outer domains, as in Fig. 3a, b, the displacements and rotations of the inner-domain and bridging-sheet components are readily apparent. When CD4 binds, the three-strand antiparallel sheet rotates by 30°, the bridging sheet forms, and the four-turn α -helix ($\alpha 1$) shifts away from the outer domain. Connecting segments, such as the link between the V1–V2 stem and the three-strand sheet or the short α -helix between the outer domain and the C-terminal strand, follow in apparent response to the movements of the secondary-structural elements. The approximate centre-of-mass displacements of inner-domain substructures (with outer domains superposed) are 28 Å, 15 Å, 12 Å and 13 Å, for the N-terminal segment, the helix $\alpha 1$, the V1–V2 stem, and the $\beta 5$ – $\beta 7$ – $\beta 25$ sheet, respectively.

The $\alpha 1$ helix is amphipathic. In the unliganded conformation it lies between the two separated β -ribbons of the bridging sheet, with its polar face exposed between them (Fig. 3a). In the CD4-bound conformation, it rotates to expose its polar face distal to the bridging sheet, which closes up over it (Fig. 3b). Intermediate positions of $\alpha 1$ would bury a set of charged residues, and intermediate conformations would therefore be unstable. This observation supports our inference, that the conformation we see represents a distinct, unliganded state, resembling the conformation present on unliganded Env trimer, rather than one of a continuum of flexible conformations selected by the crystallization conditions.

If we imagine that CD4 binding causes gp120 to shift relative to gp41, but that the latter remains essentially unaltered during this first of the series of conformational events leading to fusion, then the best frame of reference for relating the conformational change in

the gp120 core to the gp120/gp41 trimer is one in which the N and C termini of the core and the three-strand inner-domain sheet from which they emerge are fixed and all the other structural elements in gp120, including the outer domain, move with respect to them. In this frame, we have the picture shown in Fig. 3c, d. Relative to the inner-domain sheet, the outer domain rotates outwards and ‘upwards’ by 35°. The bridging sheet closes up, with a nearly 40 Å displacement of the V1–V2 stem; the site for co-receptor interaction that it forms now faces away from the assumed direction of gp41 and therefore towards the target cell.

Receptor and co-receptor binding sites

Neither the receptor (CD4) nor the co-receptor (CXCR4 or CCR5) site is properly formed in the unliganded conformation of the gp120 core (Fig. 4a). In the unliganded conformation, the CD4-binding loop projects away from the centre of the outer domain, and the $\beta 20$ – $\beta 21$ ribbon, one half of the bridging sheet, tucks in beneath it (Figs 1b and 3a). Together, the α -helices of the inner domain, the CD4-binding loop, and the $\beta 20$ – $\beta 21$ ribbon create a long, narrow cavity, lined principally with hydrophobic side chains (Fig. 4b). We identify it as a binding site for small molecules that inhibit HIV-1 entry (see below). CD4 interacts with a face of the outer domain that is partly concealed within this cavity. In the conformation stabilized by CD4 binding, the bridging sheet can close up to create the co-receptor binding surface, which is flanked by the V1–V2 and V3 loops (Figs 3b and 4a). Co-receptor specificity depends on V3, which could influence binding either directly by interacting with co-receptor or indirectly by covering part of the bridging sheet.

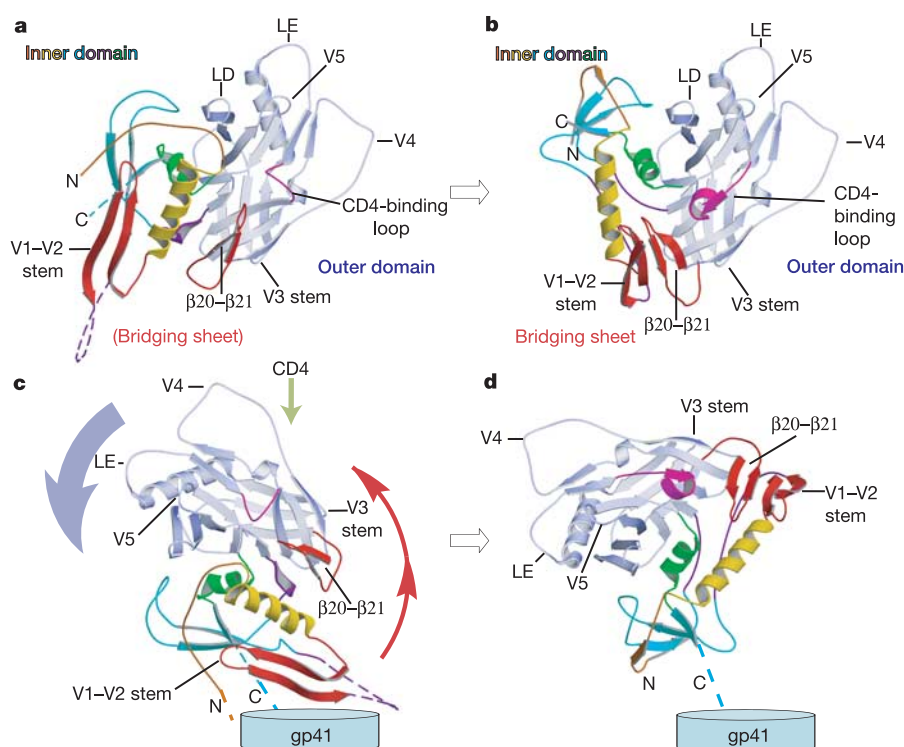


Figure 3 Conformational changes induced by CD4 binding. **a**, Unliganded SIV gp120 core structure, shown in the same colour scheme as in previous figures. The colour of the outer domain has been changed to transparent blue, to emphasize the other structural elements that undergo conformational changes upon CD4 binding. Carbohydrates are omitted for clarity. **b**, Same view of a homology model for the liganded SIV gp120 core, based on the HIV gp120 core structure in the CD4-bound conformation¹⁴. **c**, The unliganded gp120 core oriented so that the likely connection to gp41 faces ‘downwards’ as seen in the figure; gp41 is represented by the light-blue oval. The movements of the outer domain

and of the V1–V2 stem upon CD4 binding are indicated by a blue and a red arrow, respectively. CD4 enters from ‘above’ (green arrow). **d**, The gp120 core in the CD4-bound conformation, oriented to correspond to the view in **c**, when the three-strand, inner-domain β -sheet (including $\beta 5$, $\beta 7$ and $\beta 25$, in cyan) is aligned. This panel can also be viewed as the endpoint of the motions of the outer domain and V1–V2 stem indicated in **c**; the N and C termini still point downwards to gp41 (light blue), whereas the bridging sheet now faces upwards towards the target-cell membrane.

Surfaces and antigenicity

Mapping of epitopes that bind monoclonal antibodies has led to a distinction among 'neutralizing', 'non-neutralizing' and 'silent' faces of gp120 (refs 17, 25) (Fig. 5). These are thought to correspond, respectively, to epitopes on free gp120 that are exposed on the virion-associated gp120/gp41 trimer, epitopes on the free fragment that are buried on the trimer (probably at trimer interfaces or gp41 interfaces), and protein surfaces concealed by glycans. The extent of the differences between the unliganded and CD4-bound conformations now reinforces the further distinction between antibodies that recognize, stabilize, or induce one conformation and those that recognize the other. Appreciation of these two categories suggests the design of gp120 and gp160 locked in the unliganded state as selective immunogens.

Rearrangements during the transition between unliganded and CD4-bound conformations of gp120 have been invoked to explain the large negative entropy (and correspondingly large negative

enthalpy) of CD4 binding²⁶. Antibodies that recognize the receptor site on gp120 (so-called CD4bs antibodies) also have a large negative binding entropy and presumably require the same conformational transition²⁷. This entropy cost reduces antibody and receptor affinity and provides one explanation for the failure of gp120 to induce potent neutralizing antibodies against conserved features of the CD4 and co-receptor sites²⁷. A further explanation may be found in the unliganded conformation itself, which separates some of those conserved features by interspersing non-conserved elements (Fig. 4a) and which conceals others within a narrow cavity (Fig. 4b). Empirical relationships between thermodynamic parameters and conformational differences suggest that the measured entropy and enthalpy changes could either come from the ordering (upon CD4 binding) of about 100 residues or from the burial of about 10,000 Å² of molecular surface²⁶. Surface-area estimates from our unliganded SIV gp120 core structure and that of CD4-bound HIV gp120 show, after compensating for glycosylation differences and for the presence of the 17b Fab fragment, that about 7,000 Å² of surface is sequestered from solvent when CD4

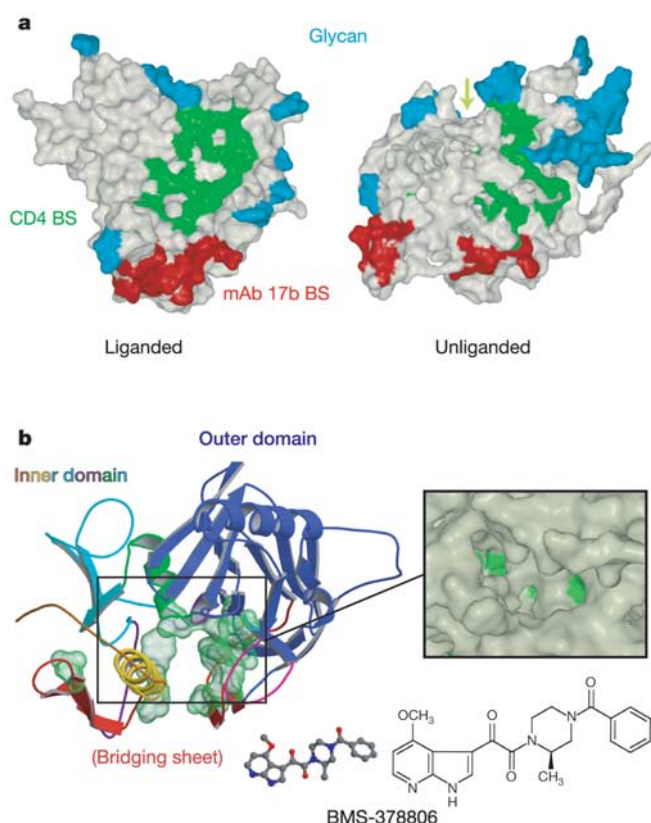


Figure 4 The binding sites (BS) of CD4 and monoclonal antibody (mAb) 17b (ref. 14), and the putative binding site of BMS-378806 (ref. 24), a small-molecule inhibitor of HIV-1 entry. **a**, Molecular surface representations of gp120 core structures in liganded and unliganded states. View as in Fig. 2a, b. Residues in direct contact with CD4 are in green; residues contacting human monoclonal antibody 17b, in red; carbohydrate, in light blue. Residues were chosen automatically from the HIV gp120 structure with the program HBPLUS⁴⁹, and were visualized with the programs LIGPLOT⁵⁰ and O. Contact criteria were quite generous, with residues defined as neighbouring if within 4.2 Å, and as hydrogen bonded if the hydrogen-to-acceptor distance was less than 3.0 Å and the donor-to-acceptor distance less than 3.6 Å. Green arrow indicates the mouth of the hydrophobic cavity shown in **b**. **b**, The unliganded gp120 core structure, as if viewed from the top in **a**. The residues lining a deep, hydrophobic cavity and corresponding to BMS-378806-selected resistance mutations in HIV gp120 are shown in surface rendering. In SIV gp120, these residues are Phe 94, Glu 95, Leu 107, Thr 272, Trp 391, Phe 398, Thr 439, Trp 440, Asn 446 and Val 485. The inset illustrates the entire cavity, with the visible resistance-mutation positions highlighted in green. The molecular formula of BMS-378806 is also shown, as well as a ball-and-stick representation on the same scale as the ribbon diagram.

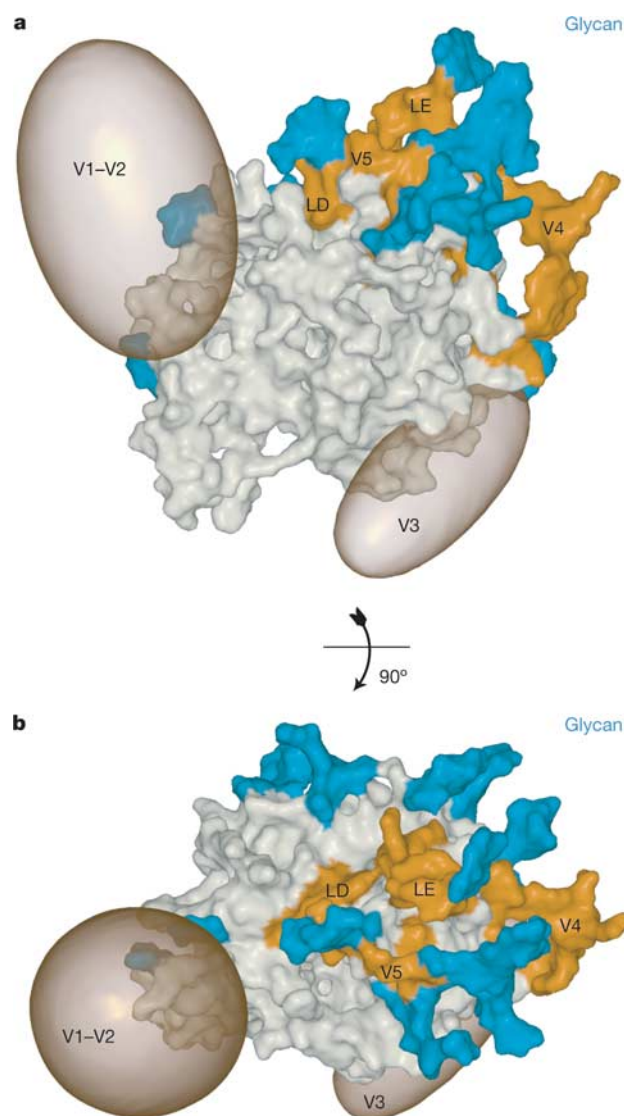


Figure 5 Surface structure of the glycosylated, unliganded gp120 core. **a**, Molecular surface of the unliganded gp120 core, viewed as in Fig. 4a. Carbohydrates are in light blue; surface loops LD and LE, and variable loops V4 and V5, in orange. Truncated variable loops V1, V2 and V3 are represented in transparent brown by mock surfaces derived from unrelated structures of comparable size. **b**, Top view of structure shown in **a**.

and gp120 associate, in reasonable agreement with the empirical prediction. There are likely also to be varying degrees of rigidification upon receptor binding, from a full disorder–order transition for residues 220–228 to reduced group motions for the inner-domain substructures.

The antibody b12 (ref. 28) is unusual in two ways: it has broad neutralizing activity and it has a markedly smaller negative entropy (as well as negative enthalpy) of binding than do other CD4bs monoclonals. Nonetheless, it probably recognizes the CD4-bound conformation, because the positions of alanine substitutions that alter its binding map to a contiguous surface on the CD4-bound gp120 structure, but map to a more distributed surface on the unliganded protein²⁹. It is possible that it relies more heavily on outer-domain contacts than do other CD4bs antibodies and hence that it requires a less complete inner-domain transition.

There are examples of monoclonal antibodies with discontinuous epitopes that map jointly to V1–V2 and V3 (refs 30, 31). These two loops project away from each other in the structure described here, and it is not likely that a single antigen-combining region of an IgG could contact both simultaneously. Models for the Env trimer based on the unliganded conformation, described below, show that the V1–V2 loop from one subunit could approach or contact the V3 loop from another subunit. We suggest that proximity of these loops from two different subunits in the trimer, rather than proximity within a single gp120, may account for the joint epitope.

A drug-binding site

A recently discovered compound, BMS-378806, and some of its

analogues, inhibit HIV-1 entry²⁴. Resistance mutations have been used to identify the residues that are likely to contact these candidate drugs^{32,33}. The sites of most of the resistance mutations lie in the deep, hydrophobic channel of unliganded gp120, as illustrated in Fig. 4b. They define a convincing binding site for molecules such as BMS-378806. The same residues do not line an obvious hydrophobic cavity in the CD4-bound structure. These observations suggest that the compounds inhibit entry by stabilizing the unliganded conformation. A group of fusion-inhibiting peptides³⁴ probably act in the same way. BMS-378806 does not inhibit SIV (ref. 24), as several of the HIV-1 resistance mutations are already present in the SIV gp120 sequence.

Conformation of gp120 in virion-associated trimers

We have created an approximate model for the unliganded trimer (Fig. 6). We have taken the structure described here as an approximation to each gp120 core on the unliganded trimer and applied the following criteria. First, the N and C termini of gp120 should point towards gp41, presumed to be centred on the three-fold axis, and roughly 'downwards' towards the viral membrane. Second, the N-linked glycans should be exposed, rather than buried at potential trimer contacts. Third, the outermost tips of the three gp120 subunits should project away from the three-fold axis, rather than cluster around it, to account for the appearance of stabilized SIV gp140 trimers in electron micrographs of negatively stained molecules³⁵. We have applied these criteria qualitatively, by visual inspection, rather than attempting to optimize a quantitative formulation. A model that satisfies them is shown in Fig. 6.

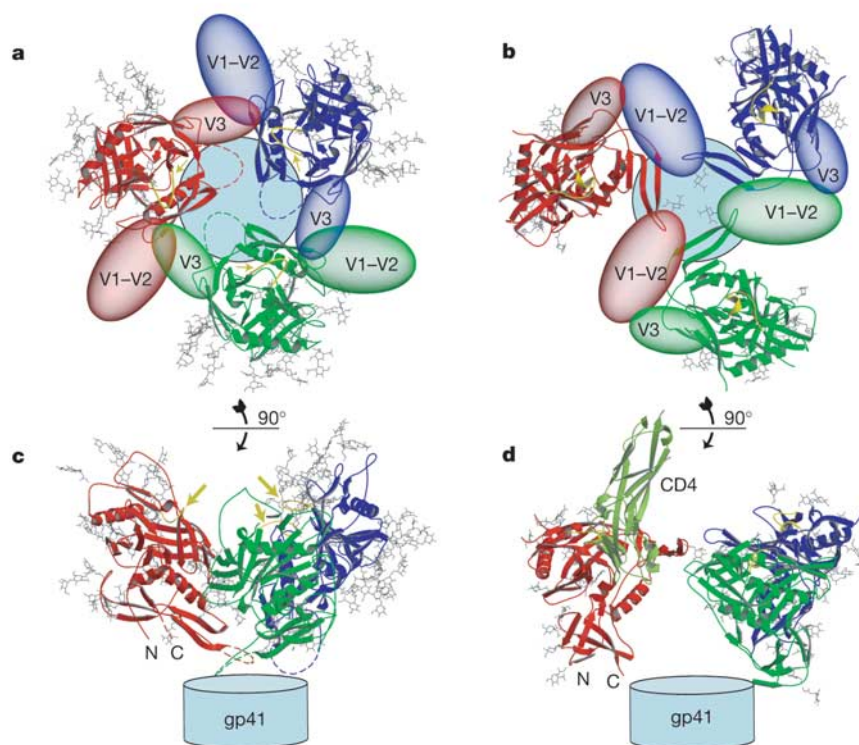


Figure 6 Proposed models for gp120/gp41 trimers in unliganded and CD4-bound conformations. **a**, A trimer in the unliganded conformation, viewed along the three-fold axis from outside the virion towards gp41. The polypeptide chain backbones are in ribbon representation; N-linked glycans are stick models; deleted V1–V2 and V3 segments are transparent balloons. The three monomers are in red, green and blue, respectively; the sugars, in grey. gp41 is shown as a cylinder in the rear. **b**, The same view of a gp120/gp41 trimer as in **a**, but in the CD4-bound conformation, generated by superposing the CD4-bound HIV gp120 structure onto the unliganded SIV gp120 subunits in panel **a**,

assuming that the three-strand, inner-domain β -sheet remains roughly in place (see Fig. 3c, d). Structural elements depicted as in **a**; CD4 omitted for clarity. **c**, 'Side' view of the same model as in **a**; the red monomer is in an orientation similar to the one in Fig. 3c. The N and C termini of the gp120 core are labelled. gp41 is shown as a cylinder at the bottom. Green arrows indicate CD4-binding loops. **d**, Side view of the same model as in **b**, with the red monomer viewed as in Fig. 3d. The first two domains of CD4 are shown in light green on only one gp120 monomer. N and C termini of the gp120 core are labelled. gp41 is shown as a cylinder at the bottom.

Although this model is only one of a range of acceptable structures, its properties are probably representative of the whole set.

In this model for an unliganded trimer, the three receptor-binding sites face a bowl-like cavity around the three-fold axis at the 'top' of the molecule (opening away from the viral membrane). The CD4-binding loop is accessible, projecting over the tip of the $\beta 20$ – $\beta 21$ ribbon. The V1–V2 stem is further recessed within the cavity, but its tip projects laterally outwards near the surface of a neighbouring gp120. A noteworthy feature of this model is the position of the disordered loop (residues from 220 to 228; dashed line in Fig. 6) connecting the V1–V2 stem to the inner-domain β -sheet. It lies near the three-fold axis, and we propose that in the unliganded trimer it contacts gp41. In the CD4-bound conformation, this loop is well ordered: the conformational change stretches it into an extended strand. Its displacement might contribute to weakening of the non-covalent gp120/gp41 association, known to accompany CD4 binding³⁶.

A further feature of our model is a potential interaction between V1–V2 of one subunit and V3 of another. In addition to accounting for the apparent joint epitope for certain antibodies, mentioned earlier, a role for both loops in stabilizing the unliganded trimer would account for a number of additional observations. TCLA virus and PBMC-adapted virus are sensitive to neutralization by soluble CD4. The main determinant of this phenotype seems to be V2 (ref. 37). Likewise, changes frequently found in CD4-independent strains are in the V1–V2 loop and within or close to HR1 of gp41 (ref. 38). These strains are also CD4-neutralization sensitive^{39,40}. Both phenotypes could result from destabilization of the unliganded conformation and hence greater ease of access by soluble CD4 to its binding site. The same may be true of the three mutations reported in an early study of T-cell line adaptation, which lie near a domain contact within a single gp120, between the N terminus of $\alpha 1$ and a loop of the outer domain, but which are also disrupted by CD4 binding⁴¹.

We have explained above that the most appropriate frame of reference for relating the CD4-bound conformation of the gp120 core to the unliganded conformation is one in which the inner-domain sheet, and the associated N and C termini, remain fixed. This assumption leads to the picture of a liganded trimer shown in Fig. 6b, d. The bridging sheet faces 'upwards', as required for interaction with the co-receptor. The orientation of the gp120 core with respect to the three-fold axis is similar to the one derived by Kwong *et al.*⁴², using criteria related to the ones we outline above. One apparent difference, but within the range of uncertainty of the two approaches, is the orientation of CD4, which approaches nearly parallel to the axis of the trimer in our model, but at about 45° from the 'side' in the Kwong *et al.* model⁴². To account for electron microscopy images of negatively stained trimers³⁵, we have also displaced the gp120 core farther from the axis than did Kwong *et al.* We suggest that most of the three-fold contacts are made through gp41 and the N- and C-terminal segments of gp120, not present in either structure.

CD4 binding and commitment to the fusion-promoting conformational transition
Even in the absence of CD4, it is likely that gp120 will fluctuate towards the CD4-bound conformation; approach of CD4 could then stabilize this excursion and ultimately fix the liganded state. An initial contact might be made at the outer-domain CD4-binding loop, which is exposed in both conformations and which moves by only about 10 Å in the transition from one to the other, as modelled in Fig. 3. Our model predicts that other parts of gp120 will shift around the CD4-binding loop with much larger excursions (for example, the tip of the V1–V2 stem moves by over 40 Å). These displacements can occur as CD4 docks. In this process, the CD4-binding loop will curl around to present three residues of extended chain for antiparallel β -sheet hydrogen-bond interaction with the C' strand at the edge of CD4 domain 1, as seen in the CD4-bound

structure¹⁴. The last of these three residues, a conserved aspartic acid, also has a bidentate salt bridge with CD4 Arg 59. The critical CD4 residue Phe 43, at the beginning of the C' strand, will then fit into a cavity between residues just distal to this aspartic acid and the edge of the bridging sheet. Reconfiguration of the loop joining the V1–V2 stem to the inner-domain sheet could, as suggested above, promote dissociation from gp41, liberating gp41 to commit to the fusion transition.

The CD4-bound structure determined by Kwong *et al.*¹⁴ includes the 17b Fab, likely to be a surrogate for the co-receptor. Is there an intermediate structure, for the state with bound CD4, but no co-receptor or antibody? A possibility, suggested by the relatively independent shifts of individual secondary-structure elements between the unliganded and CD4-bound states, is that until co-receptor (or surrogate antibody) binds, the V1–V2 stem is not docked against the $\beta 20$ – $\beta 21$ ribbon from the outer domain, and thus the bridging sheet is not properly formed. Repositioning of the V1–V2 stem, induced or fixed by the co-receptor, would then 'pull' on the loop that joins the stem to the inner-domain sheet, as outlined at the end of the preceding paragraph, and release constraints on gp41. □

Methods

Protein production and crystallization

SIV gp120 core protein was expressed in insect cells and purified by antibody affinity chromatography as described²¹. For large-scale protein production, 121 of *Trichoplusia ni* (Hi-5) cells were infected with recombinant baculoviruses at a multiplicity of infection (MOI) of 2.5. The supernatant was harvested 72 h postinfection by centrifugation, concentrated, and loaded onto an antibody 17A11 affinity column. The protein was eluted and further purified by gel filtration chromatography on Superdex 200 (Pharmacia). Fractions containing SIV gp120 core were pooled and concentrated to an absorbance of 30 at 280 nm ($A_{280} = 30$). To produce selenomethionine-substituted gp120 core protein, infected Hi-5 cells were grown in medium supplemented with 75 mg l⁻¹ selenomethionine (Sigma) and 5% FBS (Sigma), after starvation for 7 h in medium lacking methionine (JRH Biosciences). The protein was purified by the same procedure described above.

Crystals were grown in hanging drops from a mother liquor containing 15% PEG 6000, 100 mM sodium citrate, pH 5.0 and 8% PEG 400 at 20 °C. Crystals were briefly soaked in mother liquor before transfer to a solution containing 17% PEG 6000, 100 mM sodium citrate, 8% PEG 400 and 15% sucrose, in which they were flash-frozen in liquid nitrogen. Heavy atom derivatives were made by soaking crystals in mother liquor containing 1 mM of heavy atom compounds for 24 h at 20 °C.

Structure determination and trimer modelling

Crystal screening and data collection were performed at Cornell High Energy Synchrotron Source (CHESS) beamline F1 and Advanced Photon Source (APS) beamline 19ID. The best data (about 4 Å resolution) for native crystals and for crystals with bound K₂IrCl₆ or trimethyl lead acetate (TMLA) were collected at CHESS beamline F1. Data (to about 4.7 Å resolution) for SeMet crystals at peak wavelength were obtained at APS beamline 19ID. Denzo and Scalepack or HKL2000 were used for data processing⁴³. One iridium site for the K₂IrCl₆ derivative was initially identified from isomorphous and anomalous differences using the program SOLVE⁴⁴. Lead sites for the TMLA derivative and selenium sites for the selenomethionine-substituted protein were subsequently identified by difference Fourier analysis. The heavy atom sites from all the derivatives were refined together using the program SHARP⁴⁵.

Subsequent phase improvement by iterative cycles of phase combination, data sharpening, density modification, multi-crystal averaging, model building and heavily restrained refinement were performed using the CCP4 programs⁴⁶ and CNS⁴⁷. Program O⁴⁸ was used for model building. N-linked glycans from high-resolution structures deposited in the Protein Data Bank (PDB) were compiled to generate a sugar library for studying sugar conformation. Some of the fucosylated N-linked glycans from the library were fit to experimental electron density maps. Heavily restrained refinement was performed using both CNS and Refmac⁴⁶. Restraints include main-chain hydrogen-bond restraints, phi- and psi-restraints for helices, and harmonic restraints for all the glycans, main chain and side chains. Torsional simulated annealing and minimization refinement were performed in CNS. TLS refinement, followed by limited minimization with a flat bulk solvent model was implemented using the program Refmac 5. The final R_{free} and R_{work} are 38.8% and 38.5%, respectively. The relatively low data-to-parameter ratio prevented further improvement. We validated the structure by calculating averaged simulated-annealing omit maps⁴⁷, by comparing the final model with the experimentally phased map, and by calculating a model-based anomalous difference map for the SeMet sites. Full details of the structure determination and validation are given in ref. 21.

The trimer was modelled using a set of macros for O that allowed us to move one monomer freely; the two other monomers related by three-fold symmetry were generated and updated in a P3 cell. The model was judged by visual inspection based on the criteria described in main text. A trimer model for the CD-bound conformation was generated by aligning the three-strand inner-domain β -sheet as described in the caption to Fig. 6.

Received 13 October; accepted 22 December 2004; doi:10.1038/nature03327.

- Wyatt, R. & Sodroski, J. The HIV-1 envelope glycoproteins: Fusogens, antigens, and immunogens. *Science* **280**, 1884–1888 (1998).
- Allan, J. S. *et al.* Major glycoprotein antigens that induce antibodies in AIDS patients are encoded by HTLV-III. *Science* **228**, 1091–1094 (1985).
- Veronese, F. D. *et al.* Characterization of gp41 as the transmembrane protein coded by the HTLV-III/LAV envelope gene. *Science* **229**, 1402–1405 (1985).
- Center, R. J. *et al.* Oligomeric structure of the human immunodeficiency virus type 1 envelope protein on the virion surface. *J. Virol.* **76**, 7863–7867 (2002).
- Dalglish, A. G. *et al.* The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature* **312**, 763–767 (1984).
- Feng, Y., Broder, C. C., Kennedy, P. E. & Berger, E. A. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science* **272**, 872–877 (1996).
- Trkola, A. *et al.* CD4-dependent, antibody-neutralization interactions between HIV-1 and its co-receptor CCR-5. *Nature* **384**, 184–187 (1996).
- Wu, L. *et al.* CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. *Nature* **384**, 179–183 (1996).
- Sattentau, Q. J. & Moore, J. P. Conformational changes induced in the human immunodeficiency virus envelope glycoprotein by soluble CD4 binding. *J. Exp. Med.* **174**, 407–415 (1991).
- Sattentau, Q. J., Moore, J. P., Vignaux, F., Traincard, F. & Poignard, P. Conformational changes induced in the envelope glycoproteins of the human and simian immunodeficiency viruses by soluble receptor binding. *J. Virol.* **67**, 7383–7393 (1993).
- Rizzuto, C. D. *et al.* A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* **280**, 1949–1953 (1998).
- Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. Core structure of gp41 from the HIV envelope glycoprotein. *Cell* **89**, 263–273 (1997).
- Weissenhorn, W., Dessen, A., Harrison, S. C., Skehel, J. J. & Wiley, D. C. Atomic structure of the ectodomain from HIV-1 gp41. *Nature* **387**, 426–430 (1997).
- Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393**, 648–659 (1998).
- Chan, D. C., Chutkowski, C. T. & Kim, P. S. Evidence that a prominent cavity in the coiled coil of HIV type 1 gp41 is an attractive drug target. *Proc. Natl Acad. Sci. USA* **95**, 15613–15617 (1998).
- Rimsky, L. T., Shugars, D. C. & Matthews, T. J. Determinants of human immunodeficiency virus type 1 resistance to gp41-derived inhibitory peptides. *J. Virol.* **72**, 986–993 (1998).
- Wyatt, R. *et al.* The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393**, 705–711 (1998).
- Pollard, S. R., Meier, W., Chow, P., Rosa, J. J. & Wiley, D. C. CD4-binding regions of human immunodeficiency virus envelope glycoprotein gp120 defined by proteolytic digestion. *Proc. Natl Acad. Sci. USA* **88**, 11320–11324 (1991).
- Wyatt, R. *et al.* Functional and immunologic characterization of human immunodeficiency virus type 1 envelope glycoproteins containing deletions of the major variable regions. *J. Virol.* **67**, 4557–4565 (1993).
- Rud, E. W. *et al.* in *Vaccines 92: Modern Approaches to New Vaccines Including Prevention of AIDS* (eds Brown, F., Chanock, R. M., Ginsberg, H. S. & Lerner, R. A.) 229–235 (Cold Spring Harbor Laboratory, New York, 1992).
- Chen, B., Vogan, E., Gong, H. Y., Wiley, D. C. & Harrison, S. C. Determining the structure of an unliganded and fully-glycosylated SIV gp120 envelope glycoprotein. *Structure (Camb.)* (in the press).
- Kwong, P. D. *et al.* Probability analysis of variational crystallization and its application to gp120, the exterior envelope glycoprotein of type 1 human immunodeficiency virus (HIV-1). *J. Biol. Chem.* **274**, 4115–4123 (1999).
- Thali, M. *et al.* Characterization of conserved human immunodeficiency virus type 1 gp120 neutralization epitopes exposed upon gp120–CD4 binding. *J. Virol.* **67**, 3978–3988 (1993).
- Lin, P. F. *et al.* A small molecule HIV-1 inhibitor that targets the HIV-1 envelope and inhibits CD4 receptor binding. *Proc. Natl Acad. Sci. USA* **100**, 11013–11018 (2003).
- Moore, J. P. & Sodroski, J. Antibody cross-competition analysis of the human immunodeficiency virus type 1 gp120 exterior envelope glycoprotein. *J. Virol.* **70**, 1863–1872 (1996).
- Myszka, D. G. *et al.* Energetics of the HIV gp120–CD4 binding reaction. *Proc. Natl Acad. Sci. USA* **97**, 9026–9031 (2000).
- Kwong, P. D. *et al.* HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **420**, 678–682 (2002).
- Burton, D. R. *et al.* Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science* **266**, 1024–1027 (1994).
- Pantophlet, R. *et al.* Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the CD4 binding site of human immunodeficiency virus type 1 gp120. *J. Virol.* **77**, 642–658 (2003).
- Etemad-Moghadam, B. *et al.* Characterization of simian-human immunodeficiency virus envelope

- glycoprotein epitopes recognized by neutralizing antibodies from infected monkeys. *J. Virol.* **72**, 8437–8445 (1998).
- Zwick, M. B. *et al.* A novel human antibody against human immunodeficiency virus type 1 gp120 is V1, V2, and V3 loop dependent and helps delimit the epitope of the broadly neutralizing antibody immunoglobulin G1 b12. *J. Virol.* **77**, 6965–6978 (2003).
 - Guo, Q. *et al.* Biochemical and genetic characterizations of a novel human immunodeficiency virus type 1 inhibitor that blocks gp120–CD4 interactions. *J. Virol.* **77**, 10528–10536 (2003).
 - Madani, N. *et al.* Localized changes in the gp120 envelope glycoprotein confer resistance to human immunodeficiency virus entry inhibitors BMS-806 and #155. *J. Virol.* **78**, 3742–3752 (2004).
 - Ferrer, M. & Harrison, S. C. Peptide ligands to human immunodeficiency virus type 1 gp120 identified from phage display libraries. *J. Virol.* **73**, 5795–5802 (1999).
 - Chen, B. *et al.* A chimeric protein of simian immunodeficiency virus envelope glycoprotein gp140 and *Escherichia coli* aspartate transcarbamoylase. *J. Virol.* **78**, 4508–4516 (2004).
 - Moore, J. P., McKeating, J. A., Weiss, R. A. & Sattentau, Q. J. Dissociation of gp120 from HIV-1 virions induced by soluble CD4. *Science* **250**, 1139–1142 (1990).
 - Pugach, P. *et al.* The prolonged culture of human immunodeficiency virus type 1 in primary lymphocytes increases its sensitivity to neutralization by soluble CD4. *Virology* **321**, 8–22 (2004).
 - Puffer, B. A., Altamura, L. A., Pierson, T. C. & Doms, R. W. Determinants within gp120 and gp41 contribute to CD4 independence of SIV Envs. *Virology* **327**, 16–25 (2004).
 - Kolchinsky, P., Kiprilov, E. & Sodroski, J. Increased neutralization sensitivity of CD4-independent human immunodeficiency virus variants. *J. Virol.* **75**, 2041–2050 (2001).
 - Puffer, B. A. *et al.* CD4 independence of simian immunodeficiency virus Envs is associated with macrophage tropism, neutralization sensitivity, and attenuated pathogenicity. *J. Virol.* **76**, 2595–2605 (2002).
 - Turner, S. *et al.* Resistance of primary isolates of human immunodeficiency virus type 1 to neutralization by soluble CD4 is not due to lower affinity with the viral envelope glycoprotein gp120. *Proc. Natl Acad. Sci. USA* **89**, 1335–1339 (1992).
 - Kwong, P. D., Wyatt, R., Sattentau, Q. J., Sodroski, J. & Hendrickson, W. A. Oligomeric modeling and electrostatic analysis of the gp120 envelope glycoprotein of human immunodeficiency virus. *J. Virol.* **74**, 1961–1972 (2000).
 - Otwinski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
 - Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
 - de la Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276**, 472–493 (1997).
 - Collaborative Computational Project, N. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
 - Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
 - Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
 - McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793 (1994).
 - Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8**, 127–134 (1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank staff at CHESS beamline F1 and APS beamline 19ID for assistance, J. Hoxie of University of Pennsylvania, for hybridomas, and members of the Harrison/Wiley laboratory for discussion. The research was supported by a Scholar Award from the American Foundation for AIDS Research (to B.C.), by the NIH Innovation Grant Program for Approaches in HIV Vaccine Research (to S.C.H. and D.C.W.), and by an NIH HIVRAD grant (to Ellis Reinherz). S.C.H. is, and D.C.W. was, an Investigator of the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.C.H. (harrison@crystal.harvard.edu). Coordinates and structure factors have been deposited in the PDB, accession number 2BF1.

Enhanced atmospheric loss on protoplanets at the giant impact phase in the presence of oceans

Hidekazu Genda* & Yutaka Abe

Department of Earth and Planetary Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

* Present address: Department of Earth and Planetary Sciences, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8551, Japan

The atmospheric compositions of Venus and Earth differ significantly, with the venusian atmosphere containing about 50 times as much ^{36}Ar as the atmosphere on Earth¹. The different effects of the solar wind on planet-forming materials for Earth and Venus have been proposed to account for some of this difference in atmospheric composition^{2,3}, but the cause of the compositional difference has not yet been fully resolved. Here we propose that the absence or presence of an ocean at the surface of a protoplanet during the giant impact phase could have determined its subsequent atmospheric amount and composition. Using numerical simulations, we demonstrate that the presence of an ocean significantly enhances the loss of atmosphere during a giant impact owing to two effects: evaporation of the ocean, and lower shock impedance of the ocean compared to the ground. Protoplanets near Earth's orbit are expected to have had oceans, whereas those near Venus' orbit are not, and we therefore suggest that remnants of the noble-gas rich proto-atmosphere survived on Venus, but not on Earth. Our proposed mechanism explains differences in the atmospheric contents of argon, krypton and xenon on Venus and Earth, but most of the neon must have escaped from both planets' atmospheres later to yield the observed ratio of neon to argon.

According to recent planetary formation theory, the terrestrial planets were formed in two stages: the formation of several tens of Mars-sized protoplanets through accretion of planetesimals⁴, which was followed by collisions among these protoplanets^{5,6}—that is, giant impacts. We term the latter stage the stage of giant impacts.

The formation of an atmosphere by impact degassing of volatile-containing planetesimals is expected during the formation of protoplanets⁷. Simultaneously, a protoplanet gravitationally attracts the surrounding nebular gas with solar composition⁸, because protoplanetary formation (timescale $\sim 10^5 - 10^6$ yr; ref. 4) is considered to be completed before the dissipation of the surrounding nebular gas (timescale typically $\sim 10^7$ yr; ref. 9). At the stage of giant impacts (timescale $10^7 - 10^8$ yr; ref. 5), surrounding nebular gas has probably already been lost. However, the atmosphere on the protoplanet is trapped by the gravity of that protoplanet. Thus, Mars-sized protoplanets have proto-atmospheres composed of a mixture of solar and planetesimal components¹⁰. Most of the noble gases in such atmospheres are derived from the nebular gas with solar abundance, whereas most of H_2O and CO_2 are derived from planetesimals¹⁰.

When a giant impact occurs, the atmosphere near the impact site is expelled by the expansion of vapour plumes that are generated at the impact site. However, this type of direct stripping cannot remove atmosphere that lies far from the impact site—direct stripping affects only 25% of the entire atmosphere at the most¹¹. Instead, a giant impact creates a strong shock wave that travels through the planetary interior, thereby inducing a global ground motion. Such motion may expel the entire atmosphere. According to direct three-dimensional smoothed particle hydrodynamic (3D SPH) simulations of the giant impact at the escape velocity, the

velocity of the ground motion is estimated to be approximately 6 km s^{-1} at the antipode of the impact¹² and to be smaller elsewhere ($4 - 5 \text{ km s}^{-1}$ on average).

In a previous study¹¹, we performed simulations of the atmospheric motion induced by the global ground motion. We found that when a Mars-sized planet strikes an Earth-sized planet, the former loses 30% of its atmosphere, and the latter loses 10%. In other words, 90% of the atmosphere on the Earth-sized planet survives, and 70% of the atmosphere of the Mars-sized impactor is supplied to the Earth-sized planet. We also estimated that a mutual collision of planets of the same size results in a 30% atmospheric loss. Using these results, we calculated the changes in the atmospheric mass after every giant impact (Fig. 1). Figure 1 shows that the atmospheric mass after the stage of giant impacts is approximately 3–4 times the atmospheric mass of a protoplanet. This result indicates that the proto-atmospheres formed before the stage of giant impacts play an important role in the present terrestrial atmospheres¹³.

Here we focus on the effect of an ocean on the surface during the stage of giant impacts, which has not been considered in the previous studies^{11,14–16}. According to calculations made using the radiative–convective equilibrium model of an $\text{H}_2\text{O} - \text{CO}_2$ atmosphere, an ocean can form when the energy flux radiated from the planet into space (F_{pl}) is less than $\sim 300 \text{ W m}^{-2}$ (ref. 17). As the energy released by the accretion of planetesimals is negligible just after the formation of protoplanets, F_{pl} can be approximated by the net solar radiation flux, that is, $S(1 - A)/4$, where S and A are the solar radiation flux and the planetary albedo, respectively. Considering the S of the early Sun to be 70% of the present value¹⁸, that is, 960 W m^{-2} at 1 AU, F_{pl} is estimated to be 168 and 321 W m^{-2} for $A = 0.3$ at the orbits of Earth and Venus, respectively. Therefore, all the protoplanets near the Earth's orbit should have had oceans during the stage of giant impacts.

We calculate the loss fractions of the atmosphere (X_{atm}) and ocean (X_{oce}) caused by the ground motion induced by a giant

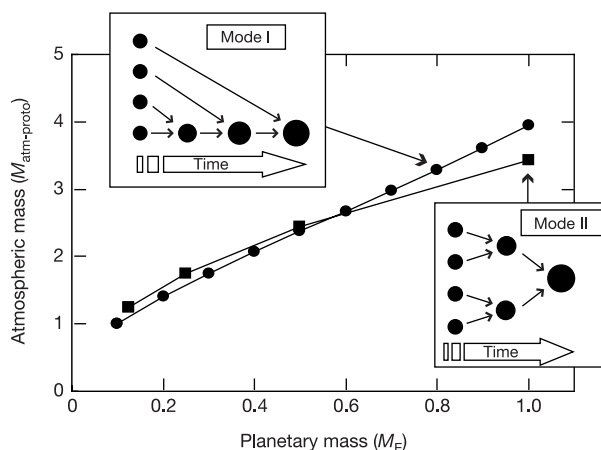


Figure 1 Changes in atmospheric mass during the stage of giant impacts. Two accretion patterns are considered. In mode I (top inset), a protoplanet grows to have the present Earth's mass (M_E) through nine giant impacts by protoplanets, each of mass $0.1 M_E$. Each protoplanet has an atmospheric mass of $M_{\text{atm-proto}}$. In mode II (bottom inset), a protoplanet grows to mass M_E through three collisions with planets of the same size; in other words, a total of eight protoplanets (each of mass $0.125 M_E$) collide with each other in the 'tournament'—each of these protoplanets has an atmospheric mass of $1.25 M_{\text{atm-proto}}$. We consider giant impacts whose collision velocities are the escape velocity. The atmospheric mass of the Earth-sized planet finally formed does not entirely depend on these accretion patterns, and is approximately 3–4 times the mass of the proto-atmosphere of a Mars-sized protoplanet.

impact, assuming a spherically symmetric motion of the atmosphere and ocean (see Methods). Figure 2 shows the calculated X_{atm} and X_{oce} as a function of the initial ground velocity (u_g). We performed simulations for various initial conditions of the atmosphere (for example, various molecular weights and temperatures), ocean, and planetary mass. It is found that the relations between X_{atm} and u_g normalized by the escape velocity are sensitive only to the initial ratio of the atmospheric mass to the ocean mass (R_{mass}), and are insensitive to other initial conditions. For a given value of u_g , X_{atm} from an ocean-covered planet is always larger than that from a planet without an ocean. There are two enhancement mechanisms for atmospheric loss. One is the vaporization of the ocean; the ground motion induces complete vaporization of the ocean, and the vaporized ocean can efficiently push out the atmosphere. The other is impedance coupling. The shock impedance of the ocean is lower than that of silicate materials, and typically higher than that of the atmosphere (see Supplementary Information). Owing to such relations of shock impedances, the velocity at the ocean–atmosphere interface becomes larger than that of the ground motion.

In the following, we consider a proto-atmosphere composed of a mixture of solar and planetesimal components. We estimate the mass of the gravitationally attracted solar component on a Mars-sized protoplanet. Assuming the minimum mass disk model of a nebula¹⁹, it is estimated to be $\sim 4 \times 10^{19}$ kg (see Supplementary Information). We assume an isothermal atmosphere, because the energy supply on the protoplanetary surface due to accretion of planetesimals has already finished during the stage of giant impacts.

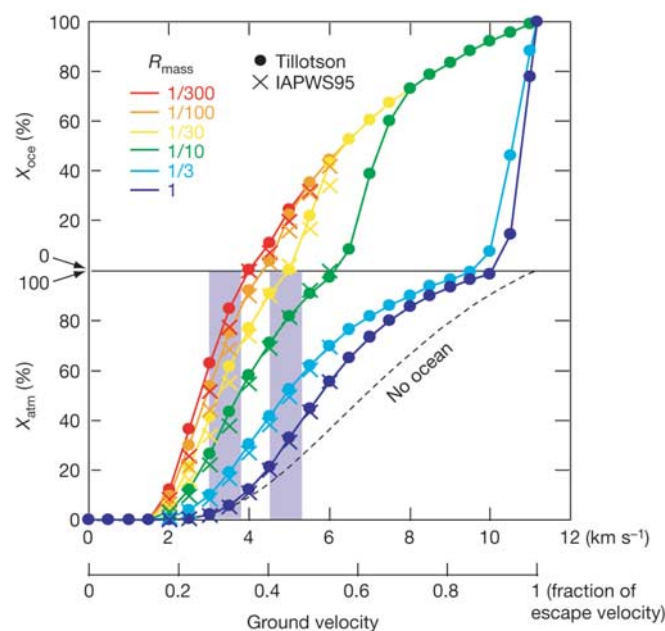


Figure 2 The loss fractions of atmosphere (X_{atm}) and ocean (X_{oce}) induced by the global ground motion with various initial ground velocities (u_g). The initial conditions of the atmosphere and ocean are described in the Methods. The results below $X_{\text{atm}} = 100\%$ imply some atmospheric loss and no oceanic loss. The results above $X_{\text{oce}} = 0\%$ imply complete atmospheric loss and some oceanic loss. In an actual giant impact, some fraction of the ocean is probably lost before the complete loss of the atmosphere. However, it is unlikely that X_{oce} is larger than X_{atm} , because the ocean initially exists below the atmosphere. Dashed curve, results without an ocean (case 1 atmosphere in figure 6 of ref. 11). Left- and right-hand vertical blue bars, range of u_g for an ocean-covered Earth-sized planet and an ocean-covered Mars-sized impactor, respectively, when the collision velocity of a giant impact is the escape velocity. u_g changes roughly linearly with the collision velocity.

As the solar abundance of ^{36}Ar is $7.6 \times 10^{-5} \text{ g g}^{-1}$ (ref. 20), this proto-atmosphere contains $\sim 3 \times 10^{15} \text{ kg}$ of ^{36}Ar , which is approximately 15 times as much as Earth's ^{36}Ar content ($2.06 \times 10^{14} \text{ kg}$), and approximately one-third of the venusian ^{36}Ar content ($1.0 \times 10^{16} \text{ kg}$)¹. The proto-atmosphere, containing a large amount of noble gases with the solar abundance, is quite different from the present Earth's atmosphere. However, such an atmosphere bears some resemblance to the present venusian atmosphere, as the venusian Ar/Kr and Kr/Xe ratios appear to be similar to the solar ones^{21,22}.

When the protoplanets are composed of planetesimals containing 1 wt% of H_2O on average¹⁰, an ocean with a maximum mass of $6 \times 10^{21} \text{ kg}$ is formed, which corresponds to $R_{\text{mass}} \approx 1/100$. Because, in reality, the ocean mass depends on the partitioning of H_2O between the proto-atmosphere and the planetary interior, here, as a reference case, we consider that $R_{\text{mass}} = 1/10$. The u_g value estimated by the 3D SPH simulations is $4\text{--}5 \text{ km s}^{-1}$ averaged over the entire surface of an Earth-sized planet. However, this value should not be directly applied to an ocean-covered planet, because the ground motion is suppressed by impedance coupling between water and rock, as compared with the free-surface case. We estimate that the decrease in u_g is approximately 25% (see Methods). Hence, we should adopt $u_g = 3\text{--}3.8 \text{ km s}^{-1}$ instead of $u_g = 4\text{--}5 \text{ km s}^{-1}$ for an Earth-sized planet covered with an ocean. From Fig. 2, X_{atm} is 30% for an Earth-sized planet with an ocean of $R_{\text{mass}} = 1/10$, while it is 10% for the case without an ocean. For a Mars-sized impactor and mutual collision of planets of the same size, X_{atm} is estimated to be 70% in the case of $R_{\text{mass}} = 1/10$, while it is 30% for the case without an ocean. In any case, no ocean escapes.

Figure 3 shows the changes in the atmospheric mass after every giant impact for an ocean-covered planet. Owing to extensive loss of the proto-atmosphere, the final atmospheric mass of an Earth-sized planet is much smaller than that without an ocean. For example, when $R_{\text{mass}} = 1/30$, it is approximately two orders of magnitude less than that on a planet without an ocean. In Fig. 3, we assume R_{mass} to be constant throughout the stage of giant impacts. However, in

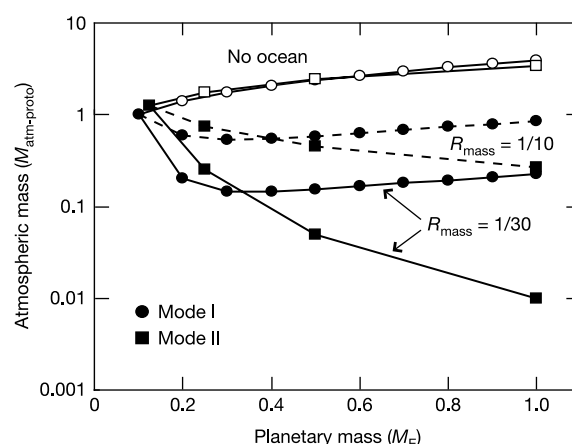


Figure 3 Changes in atmospheric mass for the cases with oceans. Filled circles and squares, results for accretion in mode I and mode II, respectively (see Fig. 1 legend). Open data points, results for the case without an ocean, which are the same as those of Fig. 1, but plotted on a logarithmic scale. We consider the cases for giant impacts whose collision velocities are the escape velocity. The amount of atmospheric loss from ocean-covered planets by each giant impact is estimated from the results shown in Fig. 2. The amount of atmosphere in the cases with an ocean becomes much smaller than that in the cases without an ocean. If impacts of ocean-covered and ocean-free planets occurred alternately, an atmospheric mass of approximately $1\text{--}2 M_{\text{atm-prot}}$ finally survives the stage of giant impacts.

reality, R_{mass} is expected to decrease after each giant impact, because a large amount of the atmosphere is lost, while the entire ocean survives. Therefore, the final amount of atmosphere on an ocean-covered planet would be smaller than that shown in Fig. 3.

As discussed before, all planets near Earth's orbit should have oceans during the stage of giant impacts. Thus, the early Earth has experienced large-scale atmospheric losses after every giant impact, but almost the entire ocean would have survived. Although ocean formation on the planets near the orbit of Venus depends on the planetary albedo, the planets inside the orbit of Venus cannot have oceans. These atmospheres are in the runaway greenhouse state. Thus, a large amount of proto-atmosphere (approximately 3–4 times the mass of a proto-atmosphere of a protoplanet) survives the giant impacts on Venus.

Just after the stage of giant impacts, a tiny amount of the proto-atmosphere (except for H_2O) remains on the Earth, while a large amount of the proto-atmosphere with a solar-like noble gas pattern remains on Venus. This can explain the large difference in Ar abundance between these planets. However, proto-atmospheres with gravitationally attracted solar components also have a large abundance of Ne. Thus, they have a Ne/Ar ratio that is about 100 times higher than those of the present atmospheres. The observed Ne/Ar ratio may be created by subsequent evolutions, such as the supply and/or erosion of the volatile components during heavy bombardment^{23–25}, and the hydrodynamic escape of hydrogen (possibly also H_2O on Venus) due to solar ultraviolet radiation^{21,22,26}.

Although we have considered a mixed proto-atmosphere of solar and planetesimal components as an example, our main result (that is, the enhancement of the atmospheric loss due to an ocean) can apply irrespective of the type of pre-existing proto-atmosphere. In any case, the presence of an ocean during planetary formation is an important factor that caused the differences between the atmosphere of Earth and of Venus. □

Methods

Motion of an atmosphere and ocean

We consider a spherically symmetric motion of an atmosphere and ocean induced by the ground motion. We ignore radiative cooling and consider no ambient nebular gas, as in the previous study¹¹. We use an ideal gas law for the EOS (equation of state) of the atmosphere. For the ocean, which is newly introduced here, we use two kinds of EOS; the Tillotson EOS²⁷ and the IAPWS95 (International Association for the Properties of Water and Steam Formulation 1995) EOS²⁸. The Tillotson EOS is widely used in the simulation of shock waves; we use the parameter sets for water²⁹. The IAPWS95 EOS is a high precision EOS specialized for H_2O , and has been used in the industrial field. We can exactly treat the vaporization of water by using the IAPWS95 EOS. Although the Tillotson EOS does not directly provide the fraction of vaporization, we can empirically estimate it from the internal energy.

As the initial conditions for an atmosphere, we consider a hydrostatically equilibrated polytropic atmosphere with given polytropic exponent (γ_a), atmospheric pressure (p_0) and temperature (T_0) at sea level, molecular weight (m_a) and specific heat ratio (γ). In Fig. 2, various values of p_0 (300, 100, 30, 10, 3 and 1 bar), $\gamma_a = 1.4$, $T_0 = 300$ K, $m_a = 2$ g mol⁻¹ and $\gamma = 1.4$ are adopted. ($p_0 = 300, 100, 30, 10, 3$ and 1 bar correspond to $R_{\text{mass}} = 1, 1/3, 1/10, 1/30, 1/100$ and $1/300$, respectively, where R_{mass} is the initial ratio of the atmospheric mass to the ocean mass.) As the initial conditions for an ocean, we consider a hydrostatically equilibrated ocean with a depth of 3 km on an Earth-sized planet. We consider a constant internal energy distribution for the Tillotson EOS with 120 J kg⁻¹, and an isentropic distribution for the IAPWS95 EOS with 300 K at sea level.

We do not solve the motion of the planetary interior induced by a giant impact. Instead, the ground motion is treated as the boundary condition of the bottom of the ocean. As in previous studies^{11,16}, we give the initial ground velocity u_g , and consider the subsequent ballistic motion of the ground—that is, the slow-down of the ground motion by gravity. We also assume that the motion ceases once the ground returns to the initial position.

Using a standard, one-dimensional, lagrangian, finite-differencing scheme³⁰, we integrate the conservational equations of the mass, momentum and energy for the atmosphere and ocean, and solve the motions of the atmosphere and ocean. We take 500 mass grids in the atmosphere, and 500 or 200 mass grids in the ocean for the Tillotson or IAPWS95 EOS, respectively. Although mixing of an ocean and atmosphere may occur due to Rayleigh–Taylor instability at the ocean–atmosphere interface, it could not be treated in the one-dimensional calculation. Nevertheless, we can estimate the upper bound of the mixing effect from the time during which the flow is subject to Rayleigh–Taylor instability. Since this duration time is less than 50 s, even if mixing occurs at the sound velocity,

the mixed region is less than 20% of the entire atmosphere and ocean. Thus, the loss fraction of the atmosphere and ocean in Fig. 2 would not be significantly affected by the instability.

Ground velocity for an ocean-covered planet

When the shock wave travelling in the planetary interior (induced by a giant impact) arrives at the ground surface, the ground surface expands and its velocity (u_g) becomes faster than the particle velocity (u_p) in the planetary interior. When the ground is covered only by the atmosphere, the ground surface can be regarded as a free surface, and u_g accelerates up to $\sim 2u_p$. When the ground is covered by an ocean, the velocity at the ground surface u_g accelerates up to $\sim 1.5u_p$ (see Supplementary Information). This difference of the acceleration of the ground surface is due to the difference in the Hugoniot curves (the relation between the particle velocity and the shock pressure) of gas and water. Therefore, the ground velocity with an ocean is slower by $\sim 25\%$ than that without an ocean, when the impact conditions are the same, and thus the particle velocities (u_p) in the planetary interior are the same in both cases.

Received 28 July 2004; accepted 12 January 2005; doi:10.1038/nature03360.

- Donahue, T. M. & Pollack, J. B. in *Venus* (eds Hunten, D., Colin, L., Donahue, T. & Moroz, V.) 1003–1036 (Univ. Arizona Press, Tucson, 1983).
- Wetherill, G. W. Solar wind origin of ^{36}Ar on Venus. *Icarus* **46**, 70–80 (1981).
- Sasaki, S. Off-disk penetration of ancient solar wind. *Icarus* **91**, 29–38 (1991).
- Kokubo, E. & Ida, S. Oligarchic growth of protoplanets. *Icarus* **131**, 171–178 (1998).
- Chambers, J. E. Making more terrestrial planets. *Icarus* **152**, 205–224 (2001).
- Kominami, J. & Ida, S. Formation of terrestrial planets in a dissipating gas disk with Jupiter and Saturn. *Icarus* **167**, 231–243 (2004).
- Abe, Y. & Matsui, T. Early evolution of the Earth: Accretion, atmosphere formation, and thermal history. *J. Geophys. Res.* **91** (suppl.), E291–E302 (1986).
- Hayashi, C., Nakazawa, K. & Mizuno, H. Earth's melting due to the blanketing effect of the primordial dense atmosphere. *Earth Planet. Sci. Lett.* **43**, 22–28 (1979).
- Storm, S. E., Edwards, S. & Skrutskie, M. F. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 837–866 (Univ. Arizona Press, Tucson, 1993).
- Abe, Y., Ohtani, E., Okuchi, T., Righter, K. & Drake, M. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 413–433 (Univ. Arizona Press, Tucson, 2000).
- Genda, H. & Abe, Y. Survival of a proto-atmosphere through the stage of giant impacts: the mechanical aspects. *Icarus* **164**, 149–162 (2003).
- Cameron, A. G. W. The giant impact revisited. *Lunar Planet. Sci. XXXIII*, 199–200 (1992).
- Melosh, H. J. The history of air. *Nature* **424**, 22–23 (2003).
- Ahrens, T. J. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 211–227 (Oxford Univ. Press, New York, 1990).
- Ahrens, T. J. Impact erosion of terrestrial planetary atmospheres. *Annu. Rev. Earth Planet. Sci.* **21**, 525–555 (1993).
- Chen, G. Q. & Ahrens, T. J. Erosion of terrestrial planet atmosphere by surface motion after a large impact. *Phys. Earth Planet. Inter.* **100**, 21–26 (1997).
- Abe, Y. Physical state of the very early Earth. *Lithos* **30**, 223–235 (1993).
- Newman, M. J. & Rood, R. T. Implications of solar evolution for the Earth's early atmosphere. *Science* **198**, 1035–1037 (1977).
- Hayashi, C. Structure of the solar nebula, growth and decay of magnetic fields and effects of magnetic and turbulent viscosities on the nebula. *Prog. Theor. Phys. Suppl.* **70**, 35–53 (1981).
- Anders, E. & Grevesse, N. Abundances of the elements: meteoritic and solar. *Geochim. Cosmochim. Acta* **53**, 197–214 (1989).
- Zahnle, K. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 1305–1338 (Univ. Arizona Press, Tucson, 1993).
- Pepin, R. O. On the origin and early evolution of terrestrial planet atmospheres and meteoritic volatiles. *Icarus* **92**, 2–79 (1991).
- Dauphas, N. & Marty, B. Inference on the nature and the mass of Earth's late veneer from noble metals and gases. *J. Geophys. Res.* **107**, doi:10.1029/2001JE001617 (2002).
- Melosh, H. J. & Vickery, A. M. Impact erosion of the primordial atmosphere of Mars. *Nature* **338**, 487–489 (1989).
- Newman, W. I., Symblasty, E. M. D., Ahrens, T. J. & Jones, E. M. Impact erosion of planetary atmospheres: Some surprising results. *Icarus* **138**, 224–240 (1999).
- Kasting, J. F. & Pollack, J. B. Loss of water from Venus. I. Hydrodynamic escape of hydrogen. *Icarus* **53**, 479–508 (1983).
- Tillotson, J. H. *Metallic Equations of State for Hypervelocity Impacts* (Report No. GA-3216, General Atomic, San Diego, California, 1962).
- Wagner, W. & Pruf, A. The IAPWS formulation 1995 for the thermodynamic properties of ordinary water substance for general and scientific use. *J. Phys. Chem. Ref. Data* **31**, 387–535 (2002).
- O'Keefe, J. D. & Ahrens, T. J. Cometary and meteorite swarm impact on planetary surfaces. *J. Geophys. Res.* **87**, 6668–6680 (1982).
- Richtmyer, R. D. & Morton, K. W. *Difference Methods for Initial-value Problems* 2nd edn (Interscience, New York, 1967).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank E. Asphaug for comments and suggestions. This work was supported by a JSPS Research Fellowship and the 21st Century Earth Science COE Program (the University of Tokyo).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.G. (genda@geo.titech.ac.jp).

Raman injection laser

Mariano Troccoli¹, Alexey Belyanin², Federico Capasso¹, Ertugrul Cubukcu¹, Deborah L. Sivco³ & Alfred Y. Cho³

¹Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

²Department of Physics, Texas A&M University, College Station, Texas 77843, USA

³Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

Stimulated Raman scattering is a nonlinear optical process that, in a broad variety of materials, enables the generation of optical gain at a frequency that is shifted from that of the incident radiation by an amount corresponding to the frequency of an internal oscillation of the material^{1,2}. This effect is the basis for a broad class of tunable sources known as Raman lasers^{2,3}. In general, these sources have only small gain ($\sim 10^{-9} \text{ cm W}^{-1}$) and therefore require external pumping with powerful lasers, which limits their applications. Here we report the realization of a semiconductor injection Raman laser designed to circumvent these limitations. The physics underlying our device differs in a fundamental way from existing Raman lasers³⁻⁸: it is based on triply resonant stimulated Raman scattering between quantum-confined states within the active region of a quantum cascade laser that serves as an internal optical pump—the device is driven electrically and no external laser pump is required. This leads to an enhancement of orders of magnitude in the Raman gain, high conversion efficiency and low threshold. Our lasers combine the advantages of nonlinear optical devices and of semiconductor injection lasers, and could lead to a new class of compact and wavelength-agile mid- and far-infrared light sources.

In these electrical injection devices the Raman shift is determined by an electronic transition between quantum-well states, known as intersubband transitions (ISTs), rather than by a phonon energy as in conventional solid state Raman lasers, and as such can be designed over a broad range.

Very large resonant nonlinear optical susceptibilities of ISTs in semiconductor quantum wells have been demonstrated since the early 1990s^{9,10}. Second harmonic generation with enhanced conversion efficiency has been reported in quantum cascade (QC) lasers using ISTs^{11,12} and interband transitions¹³. Raman lasing using IST has been theoretically discussed¹⁴ and observed experimentally in GaAs/AlGaAs double quantum wells optically pumped by a CO₂ laser^{6,7}; the Raman shift was primarily determined by a phonon resonance, anticrossed with an IST. An important step towards major performance improvements was the recent demonstration of near-infrared Raman lasers, in which ultralow threshold was achieved thanks to the use of high-quality-factor (high-Q) dielectric microsphere resonators⁸.

In a general scheme of the Raman process sketched in Fig. 1a, the internal oscillations in a medium correspond to the transition between states 1 and 2. The incident light of energy $\hbar\omega_L$ is converted into a signal of energy $\hbar\omega_S$, called Stokes radiation, where both frequencies are usually strongly detuned from other higher-lying states (that is, Δ is large compared to the broadening of level 3 in Fig. 1a) to avoid strong first-order absorption. In this case, only two-photon transitions between state 1 and 2 mediated by an intermediate virtual state may occur. As a result, the Raman gain resulting from the stimulated Raman scattering (SRS) process is of purely parametric origin as the energy of the Stokes beam is directly derived from the pump beam, that is, without intermediate storage in the medium¹.

In contrast, in resonant Raman lasing the pump and Stokes frequencies are near resonance with transitions 1–3 and 2–3 of the medium, and in turn coherently drive the transition at frequency

$\omega_{12} \approx \omega_L - \omega_S$. The triply resonant nature of this process enhances the stimulated Raman gain by many orders of magnitude with respect to the non-resonant case. In fact, the usual perturbative classification of nonlinear optical processes in powers of the electric field^{1,2} breaks down, and the effect of the fundamental radiation needs to be taken into account exactly. In this situation, the nonlinear polarization becomes comparable in magnitude to the linear terms.

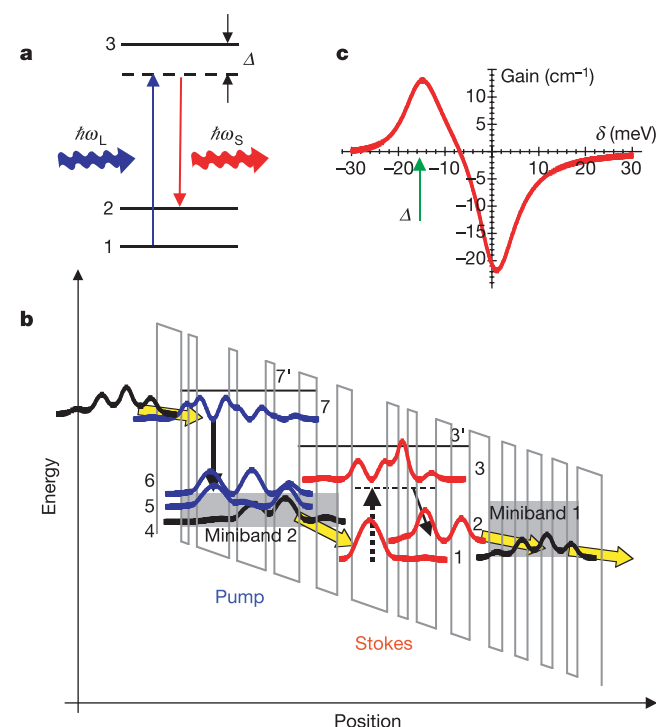


Figure 1 Diagrams showing the Raman effect and the band structure design. **a**, The Raman Stokes process. Solid and dashed lines indicate respectively real and virtual energy states. The fundamental excitation (blue), that is, the pump, is converted into a lower-energy radiation (red). Δ is the detuning of the incident radiation from the 1–3 transition resonance. **b**, Calculated conduction band structure of one period of the 30-stage Raman laser. The plot represents the potential profile along the growth direction, where the square moduli of only the most significant wavefunctions are indicated for clarity. The energy barriers (0.52 eV) are made of Al_{0.48}In_{0.52}As and the quantum wells of Ga_{0.47}In_{0.53}As. Shown are the quantum cascade laser states (4, 5, 6, 7) and Raman region states (1, 2, 3). Two higher-lying states are indicated by straight lines (3' and 7'), while the grey boxes indicate manifolds of closely spaced states (minibands 1 and 2). The layer thicknesses are (starting from the left, in nm): **4.2**, 1.3, **1.4**, 5.6, **1.4**, 4.9, **1.5**, 4.3, **3.0**, 3.6, **2.5**, 6.1, **2.0**, 1.6, **1.5**, 3.2, **2.6**, 3.2, **3.4**, **2.2**, **2.3**, **2.1**, **2.4**, 1.9, **2.5**, 1.8, **4.2**, where the barriers are indicated in boldface and the underlined layers are doped to $n = 4 \times 10^{17} \text{ cm}^{-3}$. The yellow arrows indicate the direction of electron transport. Electrons in the ground state of miniband 1 are injected by resonant tunnelling into the upper laser level (state 7) of the following period. The solid and dashed vertical arrows represent the internally generated pump laser radiation. The relevant calculated transition energies, dipole matrix elements and lifetimes are: $E_{76} = 186 \text{ meV}$, $E_{65} = 31 \text{ meV}$, $E_{54} = 39 \text{ meV}$, $E_{32} = 150 \text{ meV}$, $E_{31} = 199 \text{ meV}$, $E_{21} = 49 \text{ meV}$; $z_{32} = 1.31 \text{ nm}$, $z_{31} = 1.23 \text{ nm}$, $z_{21} = 0.7 \text{ nm}$; $\tau_{32} = 4.5 \text{ ps}$, $\tau_{31} = 1.9 \text{ ps}$, $\tau_{21} = 5.4 \text{ ps}$, $\tau_2 = 0.2 \text{ ps}$, $\tau_3 = 0.4 \text{ ps}$. The total broadenings of the transitions of the Raman section are estimated from absorption and electroluminescence data in the literature as: $\gamma_{32} = 5 \text{ meV}$, $\gamma_{31} = 5 \text{ meV}$, $\gamma_{21} = 4 \text{ meV}$. **c**, Calculated Raman gain spectrum as a function of the detuning $\delta = \omega_S - \omega_{32}$. The detuning Δ of the pump laser field from the 3–1 transition is equal to 15 meV. The drive current used in the calculation corresponds to the threshold for Stokes lasing, so that the peak gain is very close to the estimated value for the waveguide losses at the Stokes wavelength. At the peak, corresponding to the two-photon resonance $\omega_S = \omega_L - \omega_{21}$, the two-photon (Raman) term in equation (1) exceeds the linear absorption term (proportional to $n_2 - n_3$) by more than a factor of 7.

In our injection-pumped resonant Raman laser, the fundamental and the Raman radiations are both generated by electronic transitions; these transitions are between confined states in the conduction band of the very same active region of a QC laser (Fig. 1b). Raman lasing is due to the excitation of a coherent electronic polarization on the mid-infrared IST between states 1 and 2. The resonant enhancement of the third order Raman susceptibility and the intra-cavity optical pumping scheme make the process very efficient ($\sim 30\%$ power conversion). The intra-cavity pumping in each stage of the cascade overcomes the limitation of the exponential attenuation of the external pump in conventional Raman oscillators, so that the entire length of the cavity contributes to Raman lasing. Finally, the waveguide geometry ensures excellent mode overlap between the Raman and pump laser modes. The peak Raman gain coefficient per period can be estimated from the threshold condition for Stokes lasing: $\Gamma_S I_L N_P g_S = \alpha_w + \alpha_m$, where α_w and α_m are respectively the waveguide and mirror losses at the Stokes frequency ($\alpha_w + \alpha_m = 11 \pm 3 \text{ cm}^{-1}$), Γ_S is the fraction of the Stokes mode in the Raman section of the stack ($\Gamma_S \approx 0.15$), I_L is the internal pump laser intensity in the active region ($I_L = (3 \pm 0.5) \times 10^5 \text{ W cm}^{-2}$), and $N_P (= 30)$ is the number of active region periods. The Raman gain coefficient g_S is then $(8 \pm 3) \times 10^{-6} \text{ cm W}^{-1}$.

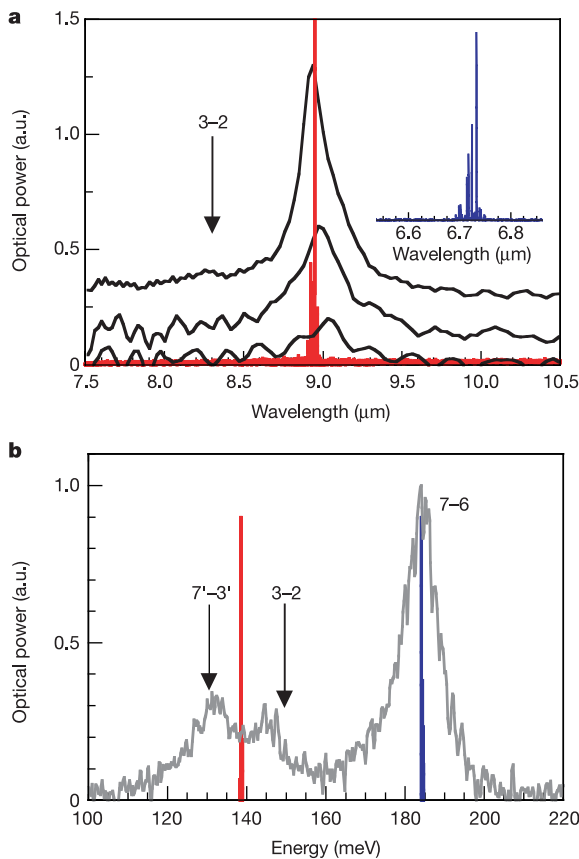


Figure 2 Spectral characteristics. **a**, Measured subthreshold emission spectra (black lines) of our QC Raman laser at currents of (from bottom to top) $I = 2.43, 2.45$ and 2.5 A, offset for clarity. The red curve is recorded above the threshold for Raman lasing ($I = 2.6$ A). Inset, fundamental laser emission spectrum at $I = 0.8$ A. The vertical arrow marks the position where the emission from the transition 3-2 would be expected if level 3 were populated by optical pumping. **b**, Comparison of electroluminescence (grey curve) and fundamental and Raman laser (blue and red curves, respectively) emission spectra. The vertical arrows indicate the energy of the transitions 3-2 and 7'-3' as calculated from the band structure in Fig. 1b. All measurements were performed in pulsed operation at a heat sink temperature of $T = 80$ K with a repetition rate of 80 kHz and a pulse width of 100 ns.

In our design of the Raman laser, the typical layout of a QC laser has been carefully modified to include within the same band structure the Λ -scheme of three ISTs designed to produce SRS. Each period of the laser multistage structure consists of an injector, a pump region and a Stokes section where Raman lasing takes place (Fig. 1b).

The pump section consists of a coupled wells vertical transition active region¹⁵, used in the design of state-of-the-art mid-infrared QC lasers¹⁶, suitably modified for optimum coupling to the Stokes section. The pump is generated across the 7-6 transition, where state 6 is depleted by resonant optical-phonon emission to level 5. The energy of the 1-3 transition (E_{13}) is designed to be detuned by 13 meV from resonant absorption of the pump. Although the detuning lowers the gain, as is shown in equation (1) below, we kept it of the order of the 1-3 transition full-width at half-maximum, $2\gamma_{31} = 10$ meV, in order to decrease optical pumping to state 3 and to be able to distinguish between Raman lasing and usual lasing due to population inversion on the 3-2 transition.

Solution of the density matrix equations and of Maxwell's equations, in a more general treatment than the one already outlined in ref. 17, gives the following expression for the signal gain at the Stokes wavelength, in units of cm^{-1} :

$$g = \text{Re} \left\{ \frac{\eta}{\gamma_{32} + |\Omega_p|^2 / (\gamma_{21} - i(\Delta - \delta)) + i\delta} \times \left[\frac{|\Omega_p|^2 (n_1 - n_3)}{(\gamma_{21} - i(\Delta - \delta))(\gamma_{31} - i\Delta)} - (n_2 - n_3) \right] \right\} \quad (1)$$

where $\Omega_p = ez_{13}E_p/\hbar$ is the Rabi frequency of the optical pump, E_p is its peak electric field, z_{13} is the dipole matrix element of the 1-3 transition and e is the electric charge, $n_{1,2,3}$ are the electron densities (per unit volume) of states 1,2,3, each γ is the total broadening (expressed as half-width at half-maximum, in frequency units) of the transition shown as a subscript, δ and Δ are the frequency detunings of the Raman and pump fields from the 3-2 and 3-1 transitions, respectively, and $\eta = 4\pi\omega_s e^2 z_{32}^2 / \hbar c$. Note that the previously defined Raman gain coefficient g_S is related to the signal gain in equation (1) through the relation $g \approx g_S I_L$, because at the peak of the gain spectrum the absorption term in equation (1) is small. Equation (1) has been derived in the so-called rotating wave approximation, valid for small frequency detunings of the pump and Stokes fields from the relevant transitions, and assuming a constant, arbitrarily strong pump field. Many-body effects have also been neglected. This is justified for doping densities in the low 10^{11} cm^{-2} range typical of our structures, leading to a frequency shift of less than 1 meV, well below the typical broadening of any transition. The Rabi frequency, for an electric field averaged over the mode profile in the active region of 14 kV cm^{-1} (that is, at pump powers of 60 mW), corresponds to an energy of ~ 2 meV, comparable to the broadening of the relevant transitions ($\gamma \approx 4-5$ meV). It would be substantially larger in the case of more powerful QC lasers that reach up to 1 W. A formula equivalent to equation (1) has been obtained¹⁸ for anti-Stokes amplification.

The gain according to equation (1) is plotted in Fig. 1c as a function of the Raman detuning δ for Δ corresponding to the band structure design. The two terms in square brackets in equation (1) demonstrate the competition between the one-photon absorption proportional to the population difference across the transition 3-2, and the nonlinear gain originating from the beating between the pump field and the off-diagonal element of the density matrix, that is, the so-called Raman coherence. In the limit when Δ becomes much larger than all linewidths, and one-photon processes are negligible, we recover from equation (1) the standard expression for the Raman gain proportional to $|E_p|^2 (n_1 - n_2) / \Delta^2$.

As is clear from equation (1), the optimal band structure design

has to meet several conditions: (1) maximize the $z_{13} \times z_{32}$ product, included in $\eta \times |\Omega_p|^2$; (2) minimize $|z_{12}|^2$ to help maximize the $z_{13} \times z_{32}$ product because of a sum-rule argument; (3) minimize state 2 lifetime τ_2 and increase n_1 to maximize Raman inversion ($n_1 \gg n_2$); and (4) optimize the detuning $\Delta = E_{31} - E_{76}$. The resulting calculated band structure is shown in Fig. 1b. n-Type doping of the injector region alone ensures a large concentration of electrons in level 1, given that the latter lies only a few meV in energy above the lowest state of miniband 1, without causing an additional broadening of the states in the Stokes section. Specific values are given in Fig. 1 legend.

The devices were based on a InGaAs/InAlAs heterostructure, grown by molecular beam epitaxy and lattice-matched to the InP substrate. Typical emission spectra from ridge waveguide devices are shown in Fig. 2a. The pump wavelength (blue curve, inset) at a heat sink temperature of 80 K is measured to be $\lambda_L = 6.7 \mu\text{m}$, corresponding very well to the predicted value of $\hbar\omega_L = 186 \text{ meV} = E_{76}$. The Stokes spectrum is peaked at a wavelength $\lambda_S = 8.9 \mu\text{m}$, corresponding to the expected value $\hbar\omega_S = \hbar\omega_L - E_{21} = 137 \text{ meV}$, and its narrowing with increased current demonstrates that it is a SRS process. The onset of Raman lasing is observed above a threshold current of 2.6 A, as indicated by the appearance of narrow cavity modes centred near λ_S . Figure 2b shows the electroluminescence spectrum of the same material processed into round mesa devices (to prevent feedback from the facets and laser action) at the same current density ($J \approx 4.5 \text{ kA cm}^{-2}$) as that in laser devices at

the threshold for Raman lasing. We can clearly identify the main peak as the pump transition E_{76} , while two low-intensity peaks appear at lower energies, neither of which corresponds to the Stokes emission. The spectral assignment of the observed features can be performed by a careful analysis of the band structure at the electric field corresponding to Raman lasing. The electric field at the threshold for Raman lasing is determined by measuring the voltage across the device and assuming that most of the voltage drop above threshold for current injection is in fact across the cladding layers, as previously demonstrated¹⁹. This leads to an estimated electric field of 51 kV cm^{-1} . The most intense electroluminescence peak is exactly centred at the pump laser photon energy, E_{76} , and the position of the Raman laser line is clearly distinguishable from the two minor peaks that we have assigned to the $7'-3'$ and $3-2$ transitions. This rules out the possibility that the Stokes line arises from conventional laser action between energy levels of the structure via electrical pumping. The upper states of these transitions are significantly detuned from the energy levels of the injector and of the active region but can still be weakly populated via hot electron effects and scattering, giving rise to the weak luminescence lines. Similarly, electroluminescence from transition $3-1$ is expected to be weak and masked by the main luminescence peak.

The laser line at $\lambda \approx 9 \mu\text{m}$ cannot arise from incoherent optical pumping of level 3, that is, our source cannot be confused with an optically pumped laser, because the $3-2$ transition is completely out of resonance with the Stokes line (Fig. 2a).

Raman lasing was reproduced in all ten tested devices. The power output characteristics of a representative device are displayed in Fig. 3a. The measured samples exhibit typically two thresholds, the first one around 1 kA cm^{-2} , for the fundamental laser emission, and a second one around 4.3 kA cm^{-2} for the Stokes emission. The two curves shown in Fig. 3a represent the laser emission turn-on of the fundamental (blue line) and Stokes (red line) radiation. The Stokes line turn-on occurs when the output laser power reaches about 40 mW, and at higher currents the Stokes power reaches about 26% of the fundamental power, where the fundamental emission starts to saturate, as commonly observed in QC lasers at injection levels of several times the threshold current. This should be compared to conversion efficiencies for solid-state Raman lasers that are usually very low ($\leq 10^{-3}$), with the exception of ultra-high-Q spherical microcavities ($\sim 20\%$) (ref. 8) and spin-flip Raman lasers in InSb ($\sim 50\%$) (ref. 2). Also shown in Fig. 3a are the emitted Stokes and fundamental output power as a function of current on a logarithmic scale, to highlight the exponential dependence of the emission below threshold as compared with the linear behaviour above threshold for both the laser lines. Given that the laser facets' reflectivity does not change significantly in the wavelength range $\lambda = 6-9 \mu\text{m}$, we can consider the ratio of the emitted powers to be equal to the ratio of the internal power densities, measured by comparing the relative intensities of the laser peaks in spectra acquired over the entire mid-infrared range without any optical filtering. From these measurements, the conversion efficiency is estimated to be about 30% at currents of 4.5 A, in good agreement with the value obtained from the power-current curves.

The thermal behaviour of the devices also shows some peculiarities. Laser action is observed up to 170 K, whereas Stokes lasing appears up to 125 K in the range of currents explored. With a threshold current density parametrized as $J = J_0 \exp(T/T_0)$, T_0 equals 97 K and 102 K for the pump and Stokes wavelengths, respectively. The emission wavelength shift with temperature is shown in Fig. 3b. The observed redshift for the pump laser is comparable to typical values for QC lasers, and amounts to $\Delta\lambda_L/\Delta T = 1.3 \text{ nm K}^{-1}$. In the case of the Raman line, there is a blueshift at increasing temperatures, $\Delta\lambda_S/\Delta T = -1.1 \text{ nm K}^{-1}$. This blueshift can be explained by recalling that the Stokes energy is given by $\hbar\omega_S = \hbar\omega_L - E_{21}$. Both $\hbar\omega_L$ and the Raman shift E_{21} are expected to decrease with increasing temperature, as is typical of ISTs²⁰.

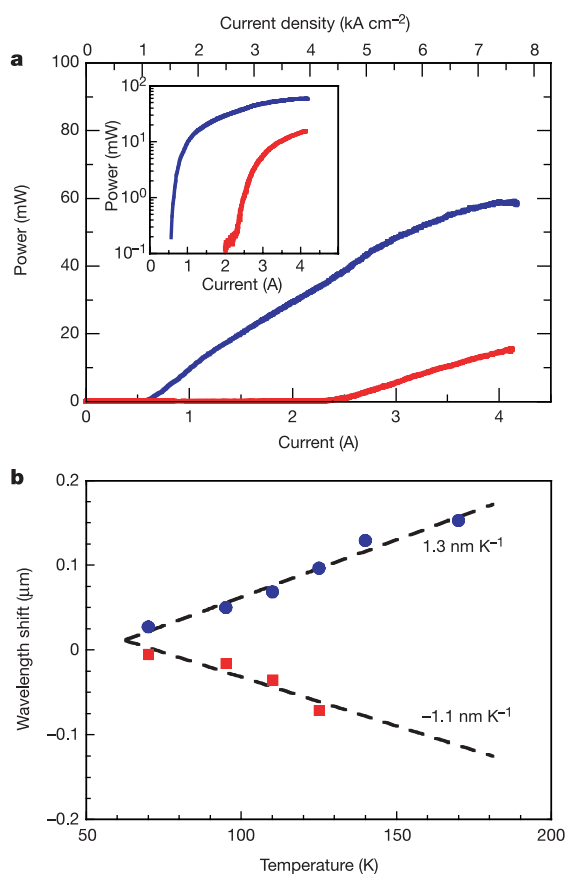


Figure 3 Power output and temperature dependence. **a**, Peak output power versus current characteristics measured for the fundamental (blue) and Stokes (red) emission at a temperature of 80 K. Measurements were performed with a 5 kHz repetition rate at a 0.05% duty cycle. Inset, experimental power-current characteristics on a logarithmic scale, showing the exponential behaviour of the emission close to threshold. **b**, Wavelength shift with temperature of the pump (blue filled circles) and Raman laser emission (red filled squares) with respect to its value at 10 K. The opposite tuning with temperature of the Stokes emission is a signature of the two-photon resonance.

However, an additional effect needs to be taken into account in our structure. E_{21} decreases at a far greater rate with increasing temperature than does $\hbar\omega_L$, because of the increased electric field required for the higher threshold current density and the large spatial separation of states 2 and 1. As a result, the 2–1 ‘diagonal’ transition shifts to the red, via the linear Stark effect, much more than does the 7–6 ‘vertical’ transition. The net effect is the observed blueshift of the Stokes emission. This behaviour therefore represents additional evidence that the emission at $9\mu\text{m}$ cannot be due to ordinary laser action. □

Methods

Device

The growth started with a $0.7\text{-}\mu\text{m}$ -thick low n -doped ($n = 5 \times 10^{16}\text{ cm}^{-3}$) GaInAs layer acting as lower waveguide core, on top of which the 30 repetitions of the active region and Raman structure periods (Fig. 1b) were grown. A $0.5\text{-}\mu\text{m}$ GaInAs layer ($n = 5 \times 10^{16}\text{ cm}^{-3}$) completes the waveguide core, on top of which an AlInAs cladding layer was grown with a total thickness of $2\text{-}\mu\text{m}$, where the first $1\text{-}\mu\text{m}$ was doped to $n = 1 \times 10^{17}\text{ cm}^{-3}$, while the rest of it was doped to $n = 5 \times 10^{17}\text{ cm}^{-3}$. The topmost layer was composed of a highly doped ($n = 4 \times 10^{18}\text{ cm}^{-3}$) $0.8\text{-}\mu\text{m}$ -thick GaInAs layer for plasmon enhanced confinement, and a final 0.1-nm -thick GaInAs contact layer Sn -doped to $n = 1 \times 10^{20}\text{ cm}^{-3}$. The material was processed into ridge waveguides 2.5 mm long and $14\text{--}20\text{-}\mu\text{m}$ wide, with a 350-nm -thick Si_3N_4 passivating layer on the lateral walls of the ridge and a $\text{Ti}(30\text{ nm})/\text{Au}(300\text{ nm})$ top contact. A non-alloyed Ge/Au contact was deposited on the back. The samples were indium-soldered on Ni/Au plated copper holders and mounted in a liquid-nitrogen flow cryostat.

Measurements

A Fourier transform infrared spectrometer was used for optical measurements, together with a calibrated room-temperature HgCdTe detector for the optical power–current characterization. To filter out only the pump laser wavelength in order to measure the optical power emitted at the Stokes frequency, a long-wavelength ($\lambda > 7.5\text{-}\mu\text{m}$) pass filter was placed along the light path.

Received 3 September 2004; accepted 4 January 2005; doi:10.1038/nature03330.

- Boyd, R. W. *Nonlinear Optics* (Academic, New York, 1992).
- Shen, Y. R. *The Principles of Nonlinear Optics* (John Wiley & Sons, Hoboken, 1984).
- Pask, H. M. The design and operation of solid-state Raman lasers. *Prog. Quant. Electron.* **27**, 3–56 (2003).
- Nishizawa, J. & Suto, K. Semiconductor Raman laser. *J. Appl. Phys.* **51**, 2429–2431 (1980).
- Grabtchikov, A. S. *et al.* All solid-state diode-pumped Raman laser with self-frequency conversion. *Appl. Phys. Lett.* **75**, 3742–3744 (1999).
- Liu, H. C. *et al.* Intersubband Raman laser. *Appl. Phys. Lett.* **78**, 3580–3582 (2001).
- Liu, H. C. *et al.* Coupled electron-phonon modes in optically pumped resonant intersubband lasers. *Phys. Rev. Lett.* **90**, 077402 (2003).
- Spillane, S. M., Kippenberg, T. J. & Vahala, L. J. Ultralow-threshold Raman laser using a spherical dielectric microcavity. *Nature* **415**, 621–623 (2002).
- Capasso, F., Sirtori, C. & Cho, A. Y. Coupled quantum well semiconductors with giant electric field tunable nonlinear optical properties in the infrared. *IEEE J. Quant. Electron.* **30**, 1313–1326 (1994).
- Rosencher, E. *et al.* Quantum engineering of optical nonlinearities. *Science* **271**, 168–173 (1996).
- Owchimikow, N. *et al.* Resonant second-order nonlinear optical processes in quantum cascade lasers. *Phys. Rev. Lett.* **90**, 043902 (2003).
- Gmachl, C. *et al.* Optimized second-harmonic generation in quantum cascade lasers. *IEEE J. Quant. Electron.* **39**, 1345–1355 (2003).
- Bengloan, J.-Y. *et al.* Intracavity sum-frequency generation in GaAs quantum cascade lasers. *Appl. Phys. Lett.* **84**, 2019–2021 (2004).
- Khurgin, J. B., Sun, G., Friedman, L. R. & Soref, R. A. Comparative analysis of optically pumped intersubband lasers and intersubband Raman oscillators. *J. Appl. Phys.* **78**, 7398–7400 (1995).
- Faist, J., Hofstetter, D., Beck, M. & Aellen, T. Bound-to-continuum and two-phonon resonance, quantum-cascade lasers for high duty cycle, high-temperature operation. *IEEE J. Quant. Electron.* **38**, 533–546 (2002).
- Capasso, F. *et al.* New frontiers in quantum cascade lasers and applications. *IEEE J. Select. Topics Quant. Electron.* **6**, 931–947 (2000).
- Belyanin, A. A., Bentley, C., Capasso, F., Kocharovskaya, O. & Scully, M. O. Inversionless lasing with self-generated driving field. *Phys. Rev. A* **64**, 013814 (2001).
- Kocharovskaya, O., Rostovtsev, Yu. V. & Imamoglu, A. Inversionless amplification in the three-level atoms with and without a hidden inversion in reservoir. *Phys. Rev. A* **58**, 649–654 (1998).
- Gmachl, C. *et al.* Dependence of the device performance on the number of stages in quantum-cascade lasers. *IEEE J. Select. Topics Quant. Electron.* **5**, 808–816 (1999).
- Helm, M. in *Intersubband Transitions in Quantum Wells: Physics and Applications I* (eds Liu, H. C. & Capasso, F.) 1–91 (Academic, London, 2000).

Acknowledgements We thank C. Gmachl for many discussions. A.B. acknowledges support from the TAMU TITF Initiative.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.T. (troccoli@deas.harvard.edu) or A. B. (belyanin@jewel.tamu.edu).

Quantum chemical calculations show that the uranium molecule U_2 has a quintuple bond

Laura Gagliardi* & Björn O. Roos*

Dipartimento di Chimica Fisica “F. Accascina”, Università degli Studi di Palermo, Viale delle Scienze – Parco d’Orleans II, I-90128 Palermo, Italy; and Department of Theoretical Chemistry, Chemical Center, POB 124, S-221 00 Lund, Sweden

* These authors contributed equally to this work

Covalent bonding is commonly described by Lewis’s theory¹, with an electron pair shared between two atoms constituting one full bond. Beginning with the valence bond description² for the hydrogen molecule, quantum chemists have further explored the fundamental nature of the chemical bond for atoms throughout the periodic table, confirming that most molecules are indeed held together by one electron pair for each bond. But more complex binding may occur when large numbers of atomic orbitals can participate in bond formation. Such behaviour is common with transition metals. When involving heavy actinide elements, metal–metal bonds might prove particularly complicated. To date, evidence for actinide–actinide bonds is restricted to the matrix-isolation³ of uranium hydrides, including $\text{H}_2\text{U} \cdot \text{UH}_2$, and the gas-phase detection⁴ and preliminary theoretical study⁵ of the uranium molecule, U_2 . Here we report quantum chemical calculations on U_2 , showing that, although the strength of the U_2 bond is comparable to that of other multiple bonds between transition metals, the bonding pattern is unique. We find that the molecule contains three electron-pair bonds and four one-electron bonds (that is, 10 bonding electrons, corresponding to a quintuple bond), and two ferromagnetically coupled electrons localized on one U atom each—so all known covalent bonding types are contributing.

Multiple chemical bonds between transition metals were unknown to inorganic chemists until the crystal structure of $\text{K}_2[\text{Re}_2\text{Cl}_8] \cdot 2\text{H}_2\text{O}$ was reported⁶ in 1965. A surprisingly short Re–Re distance of 2.24 Å was found, and assigned to a quadruple bond between the two rhenium atoms. Since then, hundreds of metal–metal multiple bonds have been characterized⁷. Here we extend the concept of metal–metal multiple bonding to the case of two actinide atoms.

The uranium atom (atomic number 92) has the ground-state electronic configuration $(5f)^3(6d)^1(7s)^2$, corresponding to a quintet ground state. However, the energy cost of unpairing the $7s$ electrons by forming hybrid orbitals is minimal, and uranium thus has in principle six electrons available with which to form chemical bonds. In a Lewis-like formalism, these electrons would combine as electron-pair bonds, giving rise to a sextuple bond between the two atoms and a singlet ground state. Such behaviour is seen with the chromium dimer⁸, where the six valence electrons reside in the $3d$ and $4s$ orbitals of the Cr atom.

But whereas the Cr atom has exactly one valence electron in each of its six valence orbitals (the five $3d$ and one $4s$ orbitals), the U atom has 16 orbitals (seven $5f$, five $6d$, one $7s$ and three $7p$) that are energetically close to one another. The bonding situation involving uranium is thus considerably more complex, given that all 16 orbitals may be considered valence orbitals available for forming the chemical bond in U_2 . This complexity makes simple inferences regarding the nature and strength of the U_2 bond impossible, despite the fact that the strength of a covalent bond depends only on the energy of the atomic orbitals on the two different centres involved, and on the overlap between the orbitals. In the case of U_2 , overlap between two $7s$ orbitals and between three out of the five $6d$

orbitals (forming one σ -type and two π -type orbitals, respectively) will be large. The $5f$ orbitals will all have smaller overlap. However, the $5f$ atomic orbital energy level is lower than that of the $6d$ and $7s$ levels, and this would favour a system that does not deviate too much from the electronic configuration of the free atom when forming the dimer.

Given this complex situation, accurate computational investigations of the bonding in U_2 call for a method that allows all valence orbitals to combine freely to form the most stable chemical bond. The complete-active-space self-consistent-field (CASSCF) method⁹ is such a method, offering maximum flexibility for describing electronic structures and capable of handling arbitrary spin. This flexibility is important because we cannot assume anything concerning the final number of paired electrons in U_2 .

CASSCF was used to generate multireference wavefunctions for subsequent multiconfigurational second-order perturbation theory calculations of the dynamic correlation energy (CASPT2)¹⁰. The effects of relativity, which are substantial for atoms as heavy as uranium¹¹, were taken into account. The calculations were performed with MOLCAS-6.0 quantum chemical software¹² (see Methods for details).

Trial studies including different sets of valence orbitals in the active space all showed that three 'normal' electron-pair bonds are formed by hybrid atomic orbitals dominated by $7s$ and $6d$ character. Double occupation of these three orbitals was thus enforced in subsequent calculations. The remaining six electrons were then allowed to freely occupy the remaining $5f$ and $6d$ orbitals while searching for the occupation resulting in the lowest energy. The most stable electronic structure was found to be a septet state (that is, the remaining six electron spins are parallel), and to have a total orbital angular momentum, Λ , equal to 11 atomic units (a.u.).

The molecular orbitals involved in forming U_2 are depicted in Fig. 1, together with the occupation number of each orbital (or pair of orbitals). As illustrated in the figure, the lowest-energy doubly occupied molecular orbital is of $7s\sigma_g$ type, corresponding to a 'typical' single σ -bond. The other two doubly occupied orbitals are degenerate, of covalent π -type and result from combinations of $6d$ and $5f$ atomic orbitals ($6d\pi_u$ in the figure). Two singly occupied orbitals, of σ -type ($6d\sigma_g$) and δ -type ($6d\delta_g$), respectively, give rise to one-electron bonds between the two atoms. Two further singly

occupied orbitals, of δ -type ($5f\delta_g$) and π -type ($5f\pi_u$), respectively, give rise to two additional (but weak) one-electron bonds. Finally, two electrons occupy what may be considered fully localized $5f\phi$ orbitals. Overall, our calculations indicate that U_2 has three strong 'normal' electron-pair bonds, two fully developed one-electron bonds, two weak one-electron bonds, and two localized electrons. The singly occupied $5f\phi$ orbitals add up to one $5f$ orbital on each atom with the electron spins parallel. In such a situation, the two electrons usually couple such that the total spin becomes zero (spin-up on one atom and spin-down on the other), because this antiferromagnetic coupling provides some additional bonding, even if small. For U_2 , however, all spins are predicted to be parallel (ferromagnetic coupling), which can be attributed to 'exchange stabilization': if all open-shell electrons have the same spin, the interaction between the non-bonding $5f$ electrons and the two one-electron bonds is energetically more favourable than the antiferromagnetic coupling of the $5f$ electrons.

The total wavefunction for the ground state can be expressed as a linear combination dominated by essentially two electronic configurations:

$$\begin{aligned} \Psi = & 0.782(7s\sigma_g)^2(6d\pi_u)^4(6d\sigma_g)^1(6d\delta_g)^1(5f\delta_g)^1(5f\pi_u)^1 \\ & \times (5f\phi_u)^1(5f\phi_g)^1 \\ & + 0.596(7s\sigma_g)^2(6d\pi_u)^4(6d\sigma_g)^1(6d\delta_g)^1(5f\delta_u)^1(5f\pi_g)^1 \\ & \times (5f\phi_u)^1(5f\phi_g)^1 + \text{small terms} \end{aligned} \quad (1)$$

All the $7s$, $6d$ and $5f$ orbitals contribute to this wavefunction. The orbitals in equation (1) have been labelled according to their angular momenta l : $l = 0$ as σ , $l = +1$ or -1 as π , $l = +2$ or -2 as δ , and $l = +3$ or -3 as ϕ , with those where l differs from zero being doubly degenerate. The total orbital angular momentum, Λ , is obtained as the sum of the angular momenta of each electron. The σ -orbitals give no contribution. The $6d\pi_u$ orbitals are occupied by four electrons (a closed shell), so the summed contribution of these electrons is also zero. The remaining orbitals are singly occupied, and the calculations show that all angular momenta have the same sign, which is in accordance with Hund's second rule. Inspection of

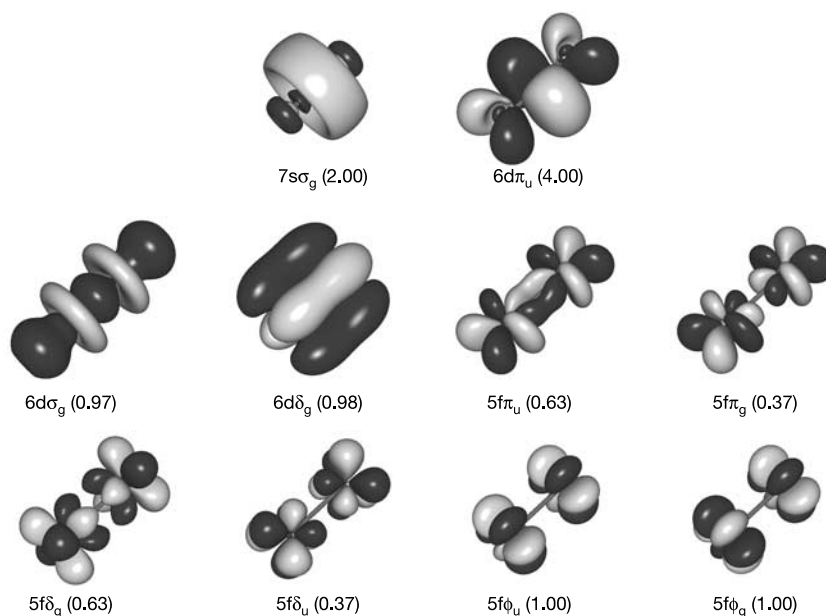


Figure 1 The active molecular orbitals forming the chemical bond between two uranium atoms. The orbital label is given below each orbital, together with the number of electrons occupying this orbital or pair of orbitals in the case of degeneracy.

the two terms in equation (1) then yields the total orbital angular momentum, $\Lambda = 0 + 2 + 2 + 1 + 3 + 3 = 11$. The calculations also show that all open shell electrons have parallel spin (Hund's first rule), resulting in a total spin angular momentum of 3, a septet state.

The U_2 chemical bond is thus more complex than any other known diatomic bond, with summation of the bonding electrons suggesting a quintuple bond. It is also unusual that strong bonding (see below) and strong ferromagnetism coexist as they do here. Moreover, its multi-radical nature (six electrons are available for binding) may support chemical bonds to a variety of ligands, such as in NU_2N or $Cl_3U_2Cl_3$. So far, the molecules OU_2O (ref. 4) and $H_2U_2H_2$ (ref. 3) have been detected.

The large number of orbitals available to the unpaired electrons permits this molecule to have the very large total orbital angular momentum of 11 a.u. However, in a heavy-atom system the combination of spin and orbital angular momenta must be considered. Our calculations show that in the case of U_2 , the total angular momentum around the molecular axis is 14 a.u. (11 a.u. from the orbitals plus 3 a.u. from the total spin). High angular momentum values have been found in lanthanide compounds, where the 4f electrons giving rise to such high values are localized and do not participate in chemical bonding. In U_2 , all electrons contribute to the bonding.

We have computed energies for a number of points in the neighbourhood of the equilibrium distance and also for two separated uranium atoms. The results for the distance with the lowest energy are presented in Table 1, giving energies where dynamic correlation is included in the calculations (CASPT2), and energies where both dynamic correlation and the effect of spin-orbit coupling are included in the calculations (CASPT2-SO). The CASPT2 level gives a dissociation energy of $40.2 \text{ kcal mol}^{-1}$, which decreases to $30.5 \text{ kcal mol}^{-1}$ when the effect of spin-orbit coupling is added. We find an equilibrium bond distance of $2.43 \text{ \AA}$ and a harmonic vibrational frequency of 265 cm^{-1} . The only experimental datum on neutral U_2 available for comparison is a dissociation energy of $52 \pm 5 \text{ kcal mol}^{-1}$ (ref. 4). Spectroscopic studies are clearly desirable to assess our results and establish the ground electronic state of U_2 , which is expected on the basis of the current work to be a $^7\Lambda_{11}$ state. However, the present computational approach has been used in a number of earlier studies of actinide compounds, yielding results in agreement with experiment both for structural properties and relative energies (see, for example, the recent study¹³ of the electronic spectrum of the UO_2 molecule). On the basis of the earlier results, we estimate that the computed U_2 bond distance is accurate to better than $0.05 \text{ \AA}$. Estimating the uncertainty in the calculated binding energy is more difficult, because a comparably complicated chemical bond has not been treated before, and because balanced treatment of the molecule and the free atoms is difficult. Nonetheless, the U_2 molecule is undoubtedly bound with an appreciable binding energy. Moreover, the fact that all low-lying energy levels correspond to the same general electronic structure assures us that our description of the binding mechanism is correct.

The first attempt⁵ to use correlated electronic structure theory to

describe the bonding in a system as complex as U_2 faced methodological and computational limitations, so the calculations were not sufficiently accurate to fully describe the bonding. Moreover, spin-orbit effects were not explicitly included. In contrast, the present results have been obtained from a CASSCF wavefunction that is based on 20 active orbitals with six active electrons, which gives much greater flexibility in the construction of the most stable electronic structure with respect to the total spin and orbital angular momenta (see Methods for details). For example, while the earlier calculations on U_2 concluded correctly that six of the valence electrons occupy the $7s\sigma_g$ and the two $6d\pi_u$ orbitals, and that most of the binding energy is contained in these three chemical bonds, the use of only six active orbitals for the remaining six electrons resulted in a $^5\Sigma_g^+$ ground state unbound with respect to dissociation. The resultant suggestion that this state of U_2 will at most be a metastable species is very different from the present findings, but can be explained by the fact that the U_2 wavefunction exhibits strong configurational mixing not accounted for in the earlier work. This situation highlights the advantages of improved methodology and computational resources available today, which have made it possible for us to carry out an accurate computational investigation of a molecule as large as U_2 and reveal its unique bonding pattern. $\square$

Methods

CASSCF calculations

The quantum chemical calculations were performed using the CASSCF method⁹. This is based on a partitioning of the molecular orbital space into three subspaces: inactive, active and external orbitals. The inactive orbitals are assumed to be doubly occupied. Remaining electrons occupy the active orbitals. The choice of these orbitals is crucial for the method. It should encompass all electronic structures that can be expected to be important for the quantum chemical problem studied. It is in principle simple to choose the active orbitals for the description of a chemical bond and its dissociation: one needs to use all orbitals that can be generated from the valence orbitals of the two atoms involved. It has been described above how the choice was made in the present case. We used a basis set of the atomic natural orbital (ANO) type that has been developed especially for relativistic calculations with the Douglas-Kroll-Hess (DKH) hamiltonian. A primitive set $26s23p17d13f5g$ was contracted to $9s8p7d5f2g$. A larger ANO basis set was constructed for the calculation of the binding energy, in which the primitive set $27s24p18d14f6g3h$ was contracted to $11s10p8d6f3g1h$.

CASPT2 and spin-orbit calculations

Dynamic correlation effects were computed using second-order perturbation theory (the CASPT2 method)¹⁰. For atoms as heavy as uranium it is necessary to include relativistic effects. They were included using the second-order DKH hamiltonian. The scalar part of this hamiltonian was used in the generation of the CASSCF wavefunction. The calculations were performed for a number of electronic states of *gerade* symmetry (after checking that states of *ungerade* symmetry were higher in energy). Spin-orbit (SO) coupling was then included by allowing the CASSCF wavefunctions to mix under the influence of the SO hamiltonian. The number of such wavefunctions was increased until the ground state energy was converged to 10^{-4} a.u. (1 a.u. = $2,625.50 \text{ kJ mol}^{-1}$). The final calculation included 16 CASSCF wavefunctions in the SO calculation. The method is described in detail in ref. 14, and references therein. It has been applied to a number of heavy atom systems with good results both for ground state properties and excited states.

Calculations were performed for a number of points around the equilibrium geometry. The binding energy was estimated from a calculation on the uranium atom, which was performed with the same basis set used for the U_2 molecule and with the valence orbitals active. In the atomic calculation, the full diatomic basis set was used to correct for the basis set superposition error. The ground state of the uranium atom is 5L , which is 17-fold degenerate. Calculations were performed for all 17 components. We used the corresponding wavefunctions as the basis in a subsequent SO calculation of the $J = 6$ ground level. The results of these calculations are presented in Table 1 and discussed in the text.

Received 8 September; accepted 2 December 2004; doi:10.1038/nature03249.

Table 1 Calculated properties of U_2

Property	CASPT2	CASPT2-SO
Total energy, $R = R_e$	-55,900.367918	-55,900.413591
Total energy, $R = R_\infty$	-55,900.299716	-55,900.364958
D_e (kcal mol ⁻¹)	40.2	30.5
R_e (Å)	2.43	2.43
ω_e (cm ⁻¹)	247	265

Total energy (atomic units) at the equilibrium bond distance, R_e , and at dissociation, R_∞ , dissociation energy, D_e , equilibrium bond distance, R_e , and harmonic vibrational frequency, ω_e , for U_2 at the CASPT2 level and at the CASPT2 plus spin-orbit coupling (CASPT2-SO) level.

- Lewis, G. N. The atom and the molecule. *J. Am. Chem. Soc.* **38**, 762–786 (1916).
- Heitler, W. & London, F. Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik. *Z. Phys.* **44**, 455–472 (1927).
- Souter, P. F., Kushto, G. P., Andrews, L. & Neurock, M. Experimental and theoretical evidence for the formation of several uranium hydride molecules. *J. Am. Chem. Soc.* **119**, 1682–1687 (1997).
- Gorokhov, L. N., Emelyanov, A. M. & Khodoev, Y. S. Mass-spectroscopic investigation of stability of gaseous molecules of U_2O_2 and U_2 . *High Temp.* **12**, 1156–1158 (1974).
- Pepper, M. & Bursten, B. E. Ab initio studies of the electronic structure of the diuranium molecule. *J. Am. Chem. Soc.* **112**, 7803–7804 (1990).
- Cotton, F. A. & Harris, C. B. The crystal and molecular structure of dipotassium octachlorodirhenate(III) dihydrate, $K_2[Re_2Cl_8] \cdot 2H_2O$. *Inorg. Chem.* **4**, 330–333 (1965).

7. Cotton, F. A. & Walton, R. A. *Multiple Bonds between Metal Atoms* (Wiley & Sons, New York, 1982).
8. Roos, B. O. The ground state potential for the chromium dimer revisited. *Collect. Czech. Chem. Commun.* **68**, 265–274 (2003).
9. Roos, B. O. in *Advances in Chemical Physics; Ab Initio Methods in Quantum Chemistry – II* Ch. 69 (ed. Lawley, K. P.) 399–445 (Wiley & Sons, Chichester, 1987).
10. Andersson, K., Malmqvist, P.-Å., Roos, B. O., Sadlej, A. J. & Wolinski, K. J. Second-order perturbation theory with a CASSCF reference function. *Phys. Chem.* **94**, 5483–5488 (1990).
11. Pyykkö, P. Relativistic effects in structural chemistry. *Chem. Rev.* **88**, 563–594 (1988).
12. Karlström, G. et al. MOLCAS: a program package for computational chemistry. *Comput. Mater. Sci.* **28**, 222–239 (2003).
13. Gagliardi, L., Heaven, M. C., Wisborg Krogh, J. & Roos, B. O. The electronic spectrum of the UO_2 molecule. *J. Am. Chem. Soc.* (in the press).
14. Roos, B. O. & Malmqvist, P.-Å. Relativistic quantum chemistry—the multiconfigurational approach. *Phys. Chem. Chem. Phys.* **6**, 2919–2927 (2004).

Acknowledgements We thank P. Pyykkö and C. J. Cramer for comments on the manuscript, P.-Å. Malmqvist and B. E. Bursten for discussions, and V. Veryazov for graphical assistance. This work was partially supported by Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR), the Swedish Research council (VR) and the Swedish Foundation for Strategic Research (SSF).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.G. (laura.gagliardi@unipa.it).

Lithospheric structure of the Rio Grande rift

David Wilson¹, Richard Aster¹, Michael West², James Ni², Steve Grand³, Wei Gao³, W. Scott Baldrige⁴, Steve Semken⁵ & Paresh Patel³

¹Department of Earth and Environmental Science and Geophysical Research Center, New Mexico Institute of Mining and Technology, Socorro, New Mexico 87801, USA

²Department of Physics, New Mexico State University, Las Cruces, New Mexico 88003, USA

³Jackson School of Geosciences, University of Texas, Austin, Texas 78712, USA

⁴Earth and Environmental Sciences Division, MS D462, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

⁵Department of Geological Sciences, Arizona State University, Tempe, Arizona 85287, USA

A high-resolution, regional passive seismic experiment^{1–6} in the Rio Grande rift region of the southwestern United States has produced new images of upper-mantle velocity structure and crust–mantle topography. Synthesizing these results with geochemical^{7–9} and other geophysical^{10–13} evidence reveals highly symmetric lower-crustal and upper-mantle lithosphere extensional deformation, suggesting a pure-shear rifting mechanism for the Rio Grande rift. Extension in the lower crust is distributed over a region four times the width of the rift's surface expression. Here we propose that the laterally distributed, pure shear extension is a combined effect of low strain rate and a regionally elevated geotherm, possibly abetted by pre-existing lithospheric structures, at the time of rift initiation. Distributed extension in the lower crust and mantle has induced less concentrated vertical mantle upwelling and less vigorous small-scale convection¹⁴ than would have arisen from more localized deformation. This lack of highly focused mantle upwelling may explain a deficit of rift-related volcanics in the Rio Grande rift compared to other major rift systems such as the Kenya rift^{15,16}.

Rifting has a profound influence on continental evolution, fundamentally controlling crustal thinning and continental breakup. The character of a continental rift depends on how lithospheric strain is accommodated. Possible strain configurations range between pure-shear and simple-shear end members. The pure-shear model is characterized by ductile deformation of the lower

crust and mantle lithosphere and predicts symmetric thinning and a symmetric lithospheric cross-section with respect to the rift axis¹⁷. The simple-shear model is characterized by strain localization along a master or a sequence of low-angle (10° to 30° dip) detachment(s) that may span the entire lithosphere¹⁸. A low-angle detachment predicts an asymmetric lithospheric cross-section, with the greatest crustal thinning laterally offset from greatest mantle lithosphere thinning.

In both the simple-shear and the pure-shear models, extensional thinning of crust and mantle lithosphere produces local upwelling of warm asthenosphere to replace thinned lithosphere. Additional heat is released by adiabatic decompression, producing partial melting and rift-associated volcanism. The emplacement of advected warmer material creates lateral temperature gradients that can induce small-scale convection^{19,20}. Small-scale convective cells may create significant additional horizontal stresses that further advance rifting, volcanism and other lithosphere-scale deformation^{14,20}.

The simple-shear and pure-shear models offer distinct topographic, heat flow, gravity anomaly, and lithospheric velocity structural predictions. Ideally, it should be possible to distinguish between the models by the topographic expression of the rift, because the asymmetry of simple-shear deformation predicts asymmetry in the flexural uplift of the rift flanks²¹, with the greatest uplift offset laterally from the surface expression of the rift. Rift-flank topography along the Rio Grande rift (RGR) (Fig. 1) is in fact relatively symmetric about the rift axis, with variations of typically less than 1–2 km between opposing rift flanks. However, the use of topography as a diagnostic tool may be significantly complicated by crustal composition and the fact that rifting has resulted in brittle deformation in the uppermost crust—expressed as a series of asymmetric grabens^{15,22}.

Similarly, the asymmetry of the simple-shear model predicts a complementary heat-flow asymmetry²³. Heat-flow measurements in the RGR region show a broad region of roughly symmetric high heat flow trending approximately along the rift axis²⁴, consistent with pure-shear extension. However, regional heat-flow values along the RGR and elsewhere may be significantly complicated by advective hydrothermal transport within the crust.

One method of assessing the existence of simple-shear detachment(s) is to image the symmetry and location of maximum crustal thinning relative to the surface expression of the rift axis. We have constructed a new image of the crust–mantle boundary using receiver functions computed from teleseismic body waves recorded by the 950-km-long, 54-station LA RISTRA experiment^{1–6}, a linear transect that crossed the RGR obliquely near 34.5° (Fig. 1). Receiver function processing isolates P-to-S converted seismic phases generated by impedance discontinuities²⁵. Discontinuity images are constructed by migration and stacking of many receiver functions with different ray paths through the crust and mantle, recorded at many stations. LA RISTRA receiver function images (Fig. 2) indicate crustal thickness ranging from 45 to 50 km beneath both the Colorado Plateau (stations NM34–UT54), and the Great Plains (TX01–NM20), with a rift-associated thinning to approximately 37 km centred beneath the RGR axis to within a few kilometres. These estimated crustal thicknesses are consistent with previous compilations of refraction surveys²⁶ across the RGR.

Using geologic constraints from previous seismic reflection work²⁴, we modelled the predicted geometry of the base of the crust (Moho) and the base of the lithosphere resulting from the amount of extension seen at the surface in the southern Albuquerque–Belen basin (16.9 km of extension over 60 km) (Figs 2, 3). The modelling technique²¹ takes into account the kinematics of lithospheric extension, the isostatic response to crust and lithospheric thinning, and the elastic response of the lithosphere (flexure). Extension by either east- or west-dipping simple shear (Fig. 2a, b) predicts offset Moho upwarping that is

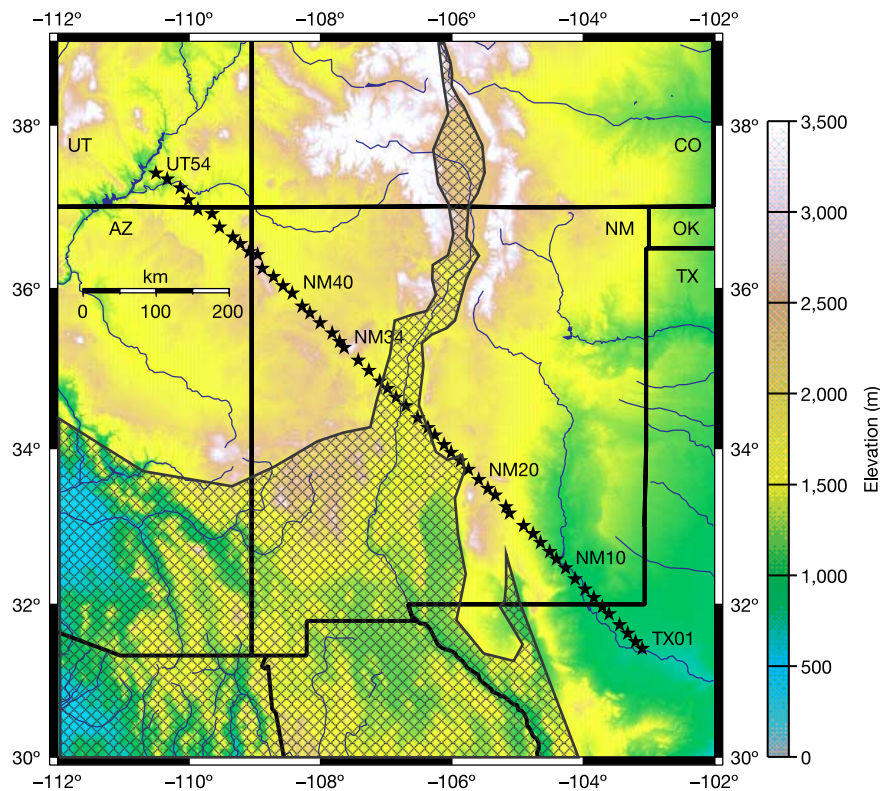


Figure 1 Elevation of the Rio Grande rift and surrounding region. The crosshatched pattern indicates the region of Cenozoic extension¹⁵. Stars indicate the broadband seismic

stations of the LA RISTRA experiment. UT, Utah; AZ, Arizona; NM, New Mexico; CO, Colorado; OK, Oklahoma; TX, Texas.

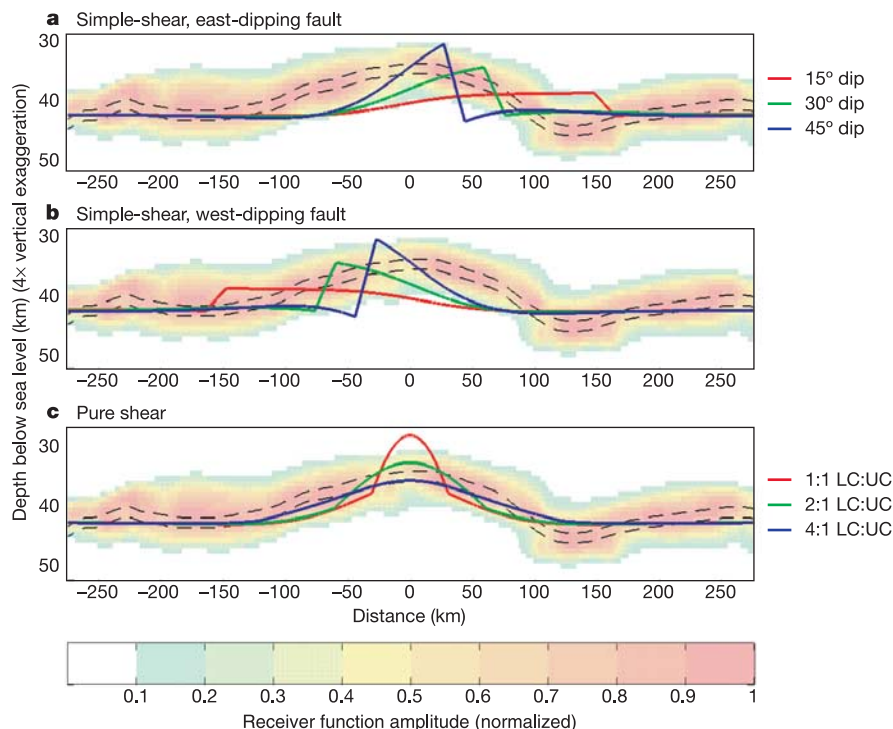


Figure 2 Modelled topography of the base of the crust. **a, b**, Predicted under simple shear extension for east-dipping faults (**a**) and west-dipping faults (**b**) with dips of 15°, 30° and 45°. **c**, Predicted from modelling pure-shear extension, with the lateral distribution of lower crustal extension relative to upper crustal extension (LC:UC) ranging from 1:1 to 4:1.

For each LC:UC ratio, the volume of crustal thinning is the same. Underlying colours show positive migrated seismic receiver function amplitudes projected onto an east–west, rift-perpendicular profile. Dashed lines represent plus and minus one standard deviation of crustal thickness (about 1.5 km).

significantly different from that observed, with the location of maximum crustal thinning offset laterally from the surface expression of the rift. The maximum misfit within 100 km of the rift axis between observed Moho depths and depths predicted by simple-shear extension for 15°, 30° and 45° east- and west-dipping faults is greater than 6 km for each case.

In contrast, modelling extension as pure shear (Fig. 2c) produces Moho upwarp that is centred on the rift axis and is highly consistent with the observed Moho geometry. Assuming that extension in the lower crust takes place over the same lateral distance (60 km) as in the upper crust (LC:UC = 1), sharp localized crustal thinning can occur, up to approximately 15 km, which is much greater and more localized than that observed. By allowing model extension in the lower crust to be distributed over a greater lateral distance we obtain more broadly distributed crustal thinning that is a much better fit to

the observed Moho upwarp. A LC:UC ratio of 4:1 (lower crustal extension over 240 km laterally) gives the smallest maximum misfit of 1.8 km within 100 km of the rift.

An observation that complements the crustal thinning profile is the lithospheric thinning profile and its rift axis symmetry. Gravity data suggest thinning of the lithosphere near the rift axis¹⁰, and previous regional teleseismic experiments^{11–13} have noted increased teleseismic delay times, suggestive of an elevated lithosphere/asthenosphere boundary. In the central RGR, the elevated lithosphere/asthenosphere boundary is centred beneath the rift, but in the southern RGR the anomaly fans out to the southwest¹³ as the RGR transitions into the Basin and Range province, a region of considerably greater Cenozoic extension. Geochemical analysis of volcanic rocks in the RGR region suggests asthenospheric ('depleted') magmatic source zones near the rift axis and lithospheric ('enriched')

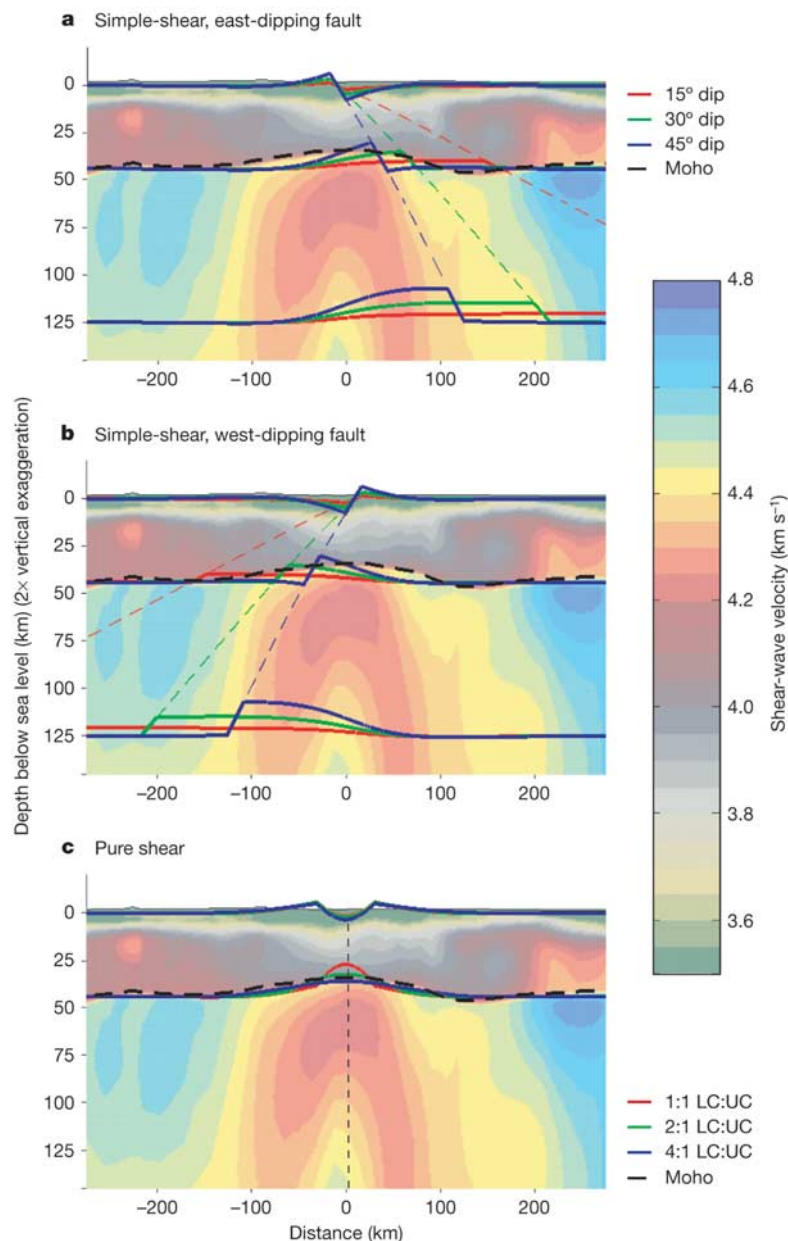


Figure 3 Lithospheric-scale cross-section in the identical projection used in Fig. 2, showing predicted surface topography and uplift of the Moho and lithosphere topography from modelling shown in Fig. 2. Underlying colours depict crust and upper-mantle shear-wave velocities determined from surface-wave phase velocity inversion⁴. Horizontal

resolution of the surface wave velocity model is 55 km in the crust and 105 km in the upper mantle. The vertical dotted line in **c** denotes the rift axis. The centring of crust and mantle low velocities on the rift axis, and the symmetry of the low velocities, indicate symmetric mantle extension about the rift axis.

source zones on the outer flanks, further supporting thinned lithosphere roughly beneath the rift axis^{7–9}.

While each of the above lines of evidence suggests that lithospheric thinning is centred approximately beneath the rift axis, this study provides the first transect view of the crust, Moho and upper mantle. Inversion of LA RISTRA surface-wave dispersion indicates that shear velocities within 100 km of the rift axis are uniformly slow throughout the crust (Fig. 3). These velocities are probably the result of increased temperatures and the possible presence of melt⁴. The symmetry of the low velocities, and inferred thermal structure, are inconsistent with highly asymmetric crustal processes. Upper-mantle shear-wave velocities (Fig. 3) reveal a transition between high velocities beneath the Great Plains and Colorado plateau and a broad region of low velocities beneath the RGR. A similar pattern of low velocities centred beneath the rift is observed in P and S body tomography². Modelled lithospheric upwarp due to simple shear extension (Fig. 3a, b) shows that the predicted location of maximum lithospheric upwarp for fault dips of 15°, 30° and 45° is offset from the rift axis by approximately 300, 180 and 100 km respectively for a reference lithospheric thickness of 125 km. Although the amount of lateral offset of maximum lithospheric upwarp depends on the reference lithospheric thickness used in modelling, simple-shear extension (Fig. 3a, b) predicts significant and observable offsets for a range of lithospheric thicknesses. Lithospheric thinning from pure shear extension, in contrast, predicts upwarp that is centred on the rift axis, and is aligned with the observed rift-centred low shear-wave velocity (Fig. 3c).

We performed numerical modelling²⁷ to compare predicted thermal and tectonic evolution of extending lithosphere by 28% over 200 km to our observed seismic imaging results. The crust is modelled as a plagioclase-rich material and the mantle is modelled as a dry olivine. Constraints on the magnitude and geometry of extension were taken from ref. 24. The modelling results show a relatively narrow zone of extension in the crust that is not matched by corresponding necking of the lithospheric mantle, where the deformation is nearly uniformly distributed. This result suggests that a pre-Cenozoic localized thermally elevated and/or compositionally weak zone was necessary to laterally concentrate strain in the mantle lithosphere underlying the RGR to the extent observed. The modelling results also show a thermal anomaly corresponding to a maximum shear-wave velocity anomaly of 2–3% (assuming $\partial \ln V_s / \partial T = -1.54 \times 10^{-4} \text{ K}^{-1}$)²⁸. RISTRA velocity inversion, where regularization probably underestimates the true variability, indicates a relative mantle shear-wave velocity anomaly of at least 8% at depths of 50–100 km (Fig. 3). This discrepancy again highlights the need for pre-existing thermal and/or compositional mantle lithosphere anomalies to explain RGR evolution, localization and present structure.

The zone of low mantle velocities is considerably wider than the minimum horizontal resolution of 150 km (or 105 km, when projected to an east–west cross-section). Thus, mantle lithosphere deformation, although spatially concentrated relative to that modelled from uniform starting conditions, is resolved to be distributed over a substantially greater lateral distance than at the surface (Fig. 3c). Major factors controlling whether lithosphere will deform by localized brittle deformation, or by distributed ductile deformation are rheology, temperature and strain rate²⁹. High temperatures and low strain rates are conducive to ductile deformation. A regionally elevated geotherm during rift initiation²⁹ is suggested by widespread regional ignimbrite volcanism around 30 Myr ago. This event, probably caused by the foundering of the subducting Farallon slab and corresponding return asthenospheric flow³⁰, may have created a zone of thermally weakened lithosphere beneath the region. Cenozoic strain rates in the central RGR are also relatively low, with only 18–28% total extension (16.9 km extension over 60 km in the southern Albuquerque–Belen basin) during the last 30 Myr^{22,24}, although strain accumulation may have been

concentrated in two distinct periods, 30–18 Myr ago and 10–5 Myr ago²². Even so, this suggests RGR formational strain rates of only 10^{-16} to 10^{-15} s^{-1} (0.56 to 1 mm yr^{-1}). This low strain rate, along with increased temperature, may have enabled ductile deformation of the lithosphere, producing the observed symmetric region of laterally distributed pure-shear deformation. Although volumes of lithospheric thinning and upwelling mantle are the same for distributed versus concentrated deformation scenarios, the observed lateral distribution of lithospheric deformation over a region that is approximately four times the width of the rift's surface expression may have resulted in less concentrated vertical mantle upwelling. The lateral temperature gradients that would have been produced by distributed extension are smaller than those expected for concentrated extension, inducing less vigorous small-scale convection and thus limiting the amount of heat delivered advectively to the shallow rift. As a result, the RGR has experienced relatively small volumes of rift-related volcanism when compared to other rift systems (for example, RGR extrusives are only 5–10% of the Kenya rift)^{15,16}. □

Received 26 February; accepted 6 December 2004; doi:10.1038/nature03297.

1. Wilson, D. *et al.* Broadband seismic background noise at temporary seismic stations observed on a regional scale in the southwestern United States. *Bull. Seismol. Soc. Am.* **92**, 3335–3341 (2002).
2. Gao, W. *et al.* Upper mantle convection beneath the central Rio Grande rift imaged by P and S wave tomography. *J. Geophys. Res.* **109**, doi:10.1029/2003JB002743 (2004).
3. Gök, R. *et al.* Shear wave splitting and mantle flow beneath LA RISTRA. *Geophys. Res. Lett.* **30**, doi:10.1029/2002GL016616 (2003).
4. West, M. *et al.* Crust and upper mantle shear wave structure of the southwest of the southwest United States: Implications for rifting and support for high elevation. *J. Geophys. Res.* **109**, doi:10.1029/2003JB002575 (2004).
5. Wilson, D. *et al.* Imaging crust and upper mantle seismic structure in the southwestern United States using teleseismic receiver functions. *Leading Edge* **22**, 232–237 (2003).
6. Wilson, D. *et al.* Seismic structure of the lithosphere in the southwestern United States using teleseismic receiver functions. *J. Geophys. Res.* (submitted).
7. Perry, F. V., Baldrige, W. S. & DePaolo, D. J. Chemical and isotopic evidence for lithospheric thinning beneath the Rio Grande rift. *Nature* **332**, 432–434 (1998).
8. Baldrige, W. S. *et al.* Middle to Late Cenozoic magmatism of the southeastern Colorado plateau and central Rio Grande rift (New Mexico and Arizona, USA)—A model for continental rifting. *Tectonophysics* **197**, 327–354 (1991).
9. McMillan, N. J. Temporal and spatial magmatic evolution of the Rio Grande rift. *New Mexico Geol. Soc. Guidebk* **49**, 107–116 (1998).
10. Cordell, L., Zorin, Y. A. & Keller, G. R. The decompensative gravity anomaly and deep structure of the region of the Rio Grande rift. *J. Geophys. Res.* **96**, 6557–6568 (1991).
11. Parker, E. C., Davis, P. M., Evans, J. R., Iyer, H. M. & Olsen, K. H. Upwarp of anomalous asthenosphere beneath the Rio Grande rift. *Nature* **312**, 354–356 (1984).
12. Davis, P. M. Continental rift structures and dynamics with reference to teleseismic studies of the Rio Grande and East-African rifts. *Tectonophysics* **197**, 309–325 (1991).
13. Slack, P. D. *et al.* The upper mantle structure of the central Rio Grande Rift region from teleseismic P- and S- wave travel time delays and attenuation. *J. Geophys. Res.* **101**, 16003–16023 (1996).
14. Mutter, J., Buck, W. R. & Zehnder, C. Convective partial melting. 1. A model for the formation of thick basaltic sequences during the initiation of spreading. *J. Geophys. Res.* **93**, 1031–1048 (1988).
15. Olsen, K. H., Baldrige, W. S. & Callender, J. F. Rio Grande rift: an overview. *Tectonophysics* **143**, 119–139 (1987).
16. Keller, G. R. *et al.* A comparative study of the Rio Grande and Kenya rifts. *Tectonophysics* **197**, 355–371 (1991).
17. McKenzie, D. Some remarks on the development of sedimentary basins. *Earth Planet. Sci. Lett.* **40**, 25–32 (1978).
18. Wernicke, B. Uniform sense normal simple shear of the continental lithosphere. *Can. J. Earth Sci.* **22**, 108–125 (1985).
19. Buck, R. W. Small-scale convection induced by passive rifting: the cause for uplift of rift shoulders. *Earth Planet. Sci. Lett.* **77**, 362–372 (1986).
20. Huismans, R. S., Podladchikov, Y. Y. & Cloetingh, S. Transition from passive to active rifting: Relative importance of asthenospheric doming and passive extension of the lithosphere. *J. Geophys. Res.* **106**, 11271–11291 (2001).
21. Weissel, J. K. & Karner, G. D. Flexural uplift of rift flanks due to mechanical unloading of the lithosphere during extension. *J. Geophys. Res.* **94**, 13919–13950 (1989).
22. Chapin, C. E. & Cather, S. M. Tectonic setting of the axial basins of the northern and central Rio Grande rift. *Geol. Soc. Am. Spec. Pap.* **291**, 5–24 (1994).
23. Buck, R. W., Martinez, F., Steckler, M. S. & Cochran, J. R. Thermal consequences of lithospheric extension: pure and simple. *Tectonics* **7**, 213–234 (1988).
24. Russell, L. R. & Snelson, S. in *Interior Rift Basins* (ed. Landon, S.) AAPG Mem. **59**, 205–258 (1994).
25. Langston, C. A. Corvallis, Oregon, crustal and upper mantle receiver structure from teleseismic P and S waves. *Bull. Seismol. Soc. Am.* **67**, 713–724 (1977).
26. Keller, G. R. & Baldrige, W. S. The Rio Grande rift: A geological and geophysical review. *Rocky Mountain Geol.* **34**, 131–148 (1999).
27. Lavie, L. L. & Buck, W. R. Half graben versus large-offset low-angle normal fault: importance of keeping cool during normal faulting. *J. Geophys. Res.* **107**, ETG 8–1–13 (2002).
28. Karato, S. Importance of anelasticity in the interpretation of seismic tomography. *Geophys. Res. Lett.* **20**, 1623–1626 (1993).

29. Morgan, P., Seager, W. & Golombek, M. Cenozoic thermal, mechanical and tectonic evolution of the Rio Grande rift. *J. Geophys. Res.* **91**, 6263–6276 (1986).
 30. Humphreys, D. Post-Laramide removal of the Farallon slab, western United States. *Geology* **23**, 987–990 (1995).

Acknowledgements We thank G. R. Keller and Roger Buck for comments. Field and data handling assistance was provided by the IRIS PASSCAL Instrument Center at the New Mexico Institute of Mining and Technology (NMT). This research was supported by NSF grants, the Los Alamos National Laboratory Institute for Geophysics and Planetary Physics, the New Mexico State University Arts and Sciences Research Center, and the NMT Geophysical Research Center. A permit is necessary to conduct geological investigations on the Navajo Nation from the Navajo Nation Minerals Department.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.W. (davew@ees.nmt.edu).

Counter-rotating microplates at the Galapagos triple junction

Emily M. Klein¹, Deborah K. Smith², Clare M. Williams² & Hans Schouten²

¹Nicholas School of the Environment and Earth Sciences, Duke University, Durham, North Carolina 27708-0227, USA

²Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

An ‘incipient’ spreading centre east of (and orthogonal to) the East Pacific Rise at 2° 40′ N has been identified as forming a portion of the northern boundary of the Galapagos microplate^{1,2}. This spreading centre was described as a slowly diverging, westward propagating rift, tapering towards the East Pacific Rise. Here we present evidence that the ‘incipient rift’ has also rifted towards the east and opens anticlockwise about a pivot at its eastern end. The ‘incipient rift’ then bounds a second microplate, north of the clockwise-rotating Galapagos microplate. The Galapagos triple junction region, in the eastern equatorial Pacific Ocean, thus consists of two counter-rotating microplates partly separated by the Hess Deep rift. Our kinematic solution for microplate motion relative to the major plates indicates that the two counter-rotating microplates may be treated as rigid blocks driven by drag on the microplates’ edges³.

The development of the Galapagos microplate (GMP) in the eastern equatorial Pacific Ocean begins with the confluence of the Cocos–Nazca, Pacific–Nazca and Pacific–Cocos spreading centres, forming a diffuse triple junction (Fig. 1b inset). Lonsdale¹ proposed that isolation of ocean crust generated at the East Pacific Rise (EPR) initiated in the south at ~1 Myr ago with the growth of a prominent seamount adjacent to the Pacific–Nazca spreading centre. This led to a zone of weakness and magma upwelling between the seamount and the adjacent EPR, forming a short, east–west-trending spreading centre. Over time, this spreading centre propagated north-eastwards, becoming the Nazca–Galapagos spreading centre. In the current plate configuration, this plate boundary widens and deepens as it approaches the southern bounding scarps of the Galapagos gore, near 101° W, in what is called the Dietz Deep.

In contrast to the relatively well-understood southern boundaries of the GMP, the configuration of its northern boundary has not been fully elucidated. Early studies suggesting that the northern boundary consisted of the westward extension of the Cocos–Nazca spreading centre^{4,5} were later refuted by the finding that the Cocos–Nazca spreading centre does not link to the EPR, but rather terminates ~50 km east of it at the Hess Deep^{6,7}. Subsequent surveys further north, however, identified the ‘incipient rift’ (IR), a magmatically active, newly forming, spreading centre east of and

orthogonal to the EPR at ~2° 40′ N (refs 1, 2, 8), forming at least a portion of the northern boundary of the GMP. A study of the western portion of the IR suggested that it consists of a slowly diverging (~15 mm yr⁻¹), westward propagating, spreading centre between the Cocos and Galapagos plates, which leaves a wedge-shaped gore in its wake².

In August 2002, aboard the R/V *Melville*, we mapped and sampled a broad region centred on the IR, including its intersection with the EPR and extending ~140 km to the east. Various geological, geophysical and rock sampling tools were used, including SeaBeam2000 bathymetry and side-scan (amplitude) coverage; towed magnetometer; bottom photography (14 camera tows using the WHOI Towed Camera System); water column hydrothermal surveying⁹; and rock sampling¹⁰. Here we present findings based primarily on our bathymetric, side-scan, magnetic and camera tow results.

The east–west-trending IR is shallowest (~2,900 m below sea level) near its intersection with the EPR (at ~2° 40′ N), and deepens progressively eastward (>3,500 m), exceeding average depths of ambient sea floor to its north (Fig. 1a). The loci of greatest depths along the eastern half of the IR form a sinuous trough that trends ~100° to the southeast. SeaBeam amplitudes for the IR indicate high reflectivity consistent with relatively sparse sediment cover compared to adjacent sea floor (Fig. 2a). Along the eastern portion of the IR, reflection amplitudes delineate an eastward-narrowing, highly reflective ‘wedge’ that coincides with the location of the bathymetric trough. The apex of this wedge clearly cuts north–south-oriented abyssal hills and then dies out in the vicinity of 101° 30′ W (Fig. 2a).

The photographic and magnetic data support the idea that the IR, including the eastern reflective trough, is magmatically active. Photographs of the crust along the IR show sparsely sedimented lavas, often with delicate ornamentation and basaltic glass (recovered by dredging), with local in-filling of sediment between pillows (Fig. 2a–c). In a number of photographs, particularly along the eastern portion of the IR, lavas appear to emanate from local fissures with east–west orientations (for example, Fig. 2b). Previous studies of northern EPR crust show that within a few kilometres of the ridge axis, undisturbed lavas generally become buried under a thick pile of sediment¹¹. Although not as fresh as lavas observed at recent EPR axial eruptions, the lavas photographed along the eastern IR appear significantly younger than the calculated crustal age of ~400–700 kyr (based on the EPR spreading rate), and therefore probably erupted *in situ*. In the absence of radiometric age dates, we speculate that these lavas are a few thousand years old, based on age-dating studies performed on lavas photographed in place elsewhere¹¹. This suggests that although the IR is magmatically active along its length, it may be only sporadically so, interspersed with periods of magmatic inactivity and sediment burial, consistent with its slow spreading rate.

Magnetic anomaly profiles run over the IR also support recent magmatic activity (Fig. 1b). Using magnetic field data (collected on our own cruise and on previous cruises), we calculated crustal magnetization, taking into account bathymetry and assuming a constant source thickness of 0.5 km (refs 12, 13). A crustal magnetization high is centred over the IR at ~101° 43′ N, which can be explained by a thicker magnetic source layer, younger and therefore more highly magnetized rocks (for example, erupted <10 kyr ago), and/or lavas with higher Fe or Ti contents than the surrounding region. The higher iron contents of basalts (13–16 wt% Fe₂O₃)¹⁰ dredged from this region can explain only half the amplitude of the magnetization¹⁴. We suggest that the other half is caused by the high magnetization of young lavas that record the high geomagnetic field intensity of the past 10 kyr (ref. 15).

The fact that the eastern portion of the reflective IR wedge cuts north–south-oriented abyssal hills, combined with the side-scan, magnetic and camera-tow data suggesting recent magmatism along

this wedge, raises the possibility that the IR is opening like a small spenochasm¹⁶ (a wedge-shaped opening) about a pivot located near the eastern tip of the wedge. Thus, the IR can be viewed as a lozenge-shaped feature, with the western taper reflecting westward growth of the IR at the EPR triple junction¹, and the eastern taper indicating opening about a stable pivot at the eastern tip (Figs 1 and 2).

If the IR opens about a pivot at its eastern end, lithosphere immediately south of the IR must rotate anticlockwise about this pivot (Fig. 3). This anticlockwise rotation contrasts with the known clockwise rotation of the southern portion of the GMP, south of the Hess Deep rift³. It follows that what was previously considered one coherent microplate, must, in fact, form two separate microplates (Fig. 3): the northern portion of the GMP (NGMP), rotating anticlockwise, and the remaining portion of the GMP to the south, rotating clockwise.

We estimate the kinematics of the GMP and NGMP by drawing upon the concepts of edge-driven microplate mechanisms³. In this approach, the rotation of a microplate is driven by a shear couple between a pair of bounding plates moving in opposite directions. The two points of coupling between the microplate and bounding plates are represented by two instantaneous relative rotation axes (IRRA). In edge-driven microplate systems like the Easter¹⁷ and Juan Fernandez¹⁸ microplates, these axes commonly lie ahead of the tips of the microplate bounding rifts.

We identify three likely IRRA locations—labelled NG-C, G-N and NG-G in Fig. 3a—at the tips of three of the rifts that bound the microplates, that is, at the eastern end of the IR, at the Hess Deep

and at the Dietz Deep. Cocos–Nazca–Pacific motion¹⁹ and distances between IRRAs scale the relative rotation rates about those axes. The three axes and major plate motions (Fig. 3a inset) determine a solution for the instantaneous motion of the two microplates (Fig. 3b) that satisfies the triple junction geometries at $2^{\circ}40'N^1$ and $1^{\circ}10'N^2$. Details of the kinematic modelling are provided in Supplementary Information. The solution in Fig. 3b, which has NG-C, G-N and NG-G IRRAs respectively located on the NGMP–Cocos, GMP–Nazca, and NGMP–GMP boundaries, suggests that the counter-rotating microplates can be modelled as rigid blocks and that their rotation could be driven by drag on the microplates' edges rather than by shear flow of mantle underneath³.

The maximum width of the lozenge shape of the IR marks the location and time (0.5 Myr ago) when the IR was established adjacent to the EPR. Since that time, it has propagated eastward at a rate that we are not able to determine (that is, instantaneously as a crack or more slowly). Estimated spreading rates predict that since its initiation at 0.5 Myr ago, a maximum of ~ 7.5 km of new IR-generated crust should have formed, a distance that agrees well with the maximum width of the lozenge (Fig. 1). The IR sea floor at this location would also have been subjected to $\sim 6.5^{\circ}$ of rotation, an angle that is probably too small to be resolved in the form of deformed fabric of the sea floor.

Considered in the evolutionary history of the GMP region, we speculate that the eastern tip of the IR or Dietz Deep may ultimately break through the bounding walls of the Hess Deep gore to connect with the Cocos–Nazca spreading centre. If this happens, the driving torque for the microplate will cease, and the microplates will stop

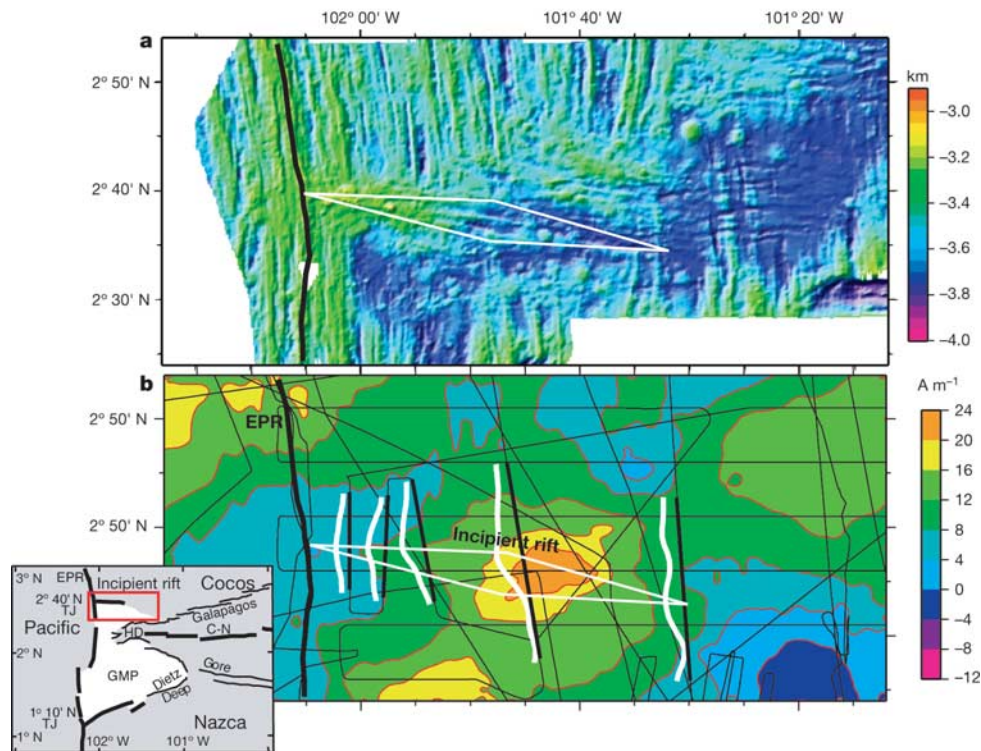


Figure 1 Bathymetric and magnetic data collected in the vicinity of the IR. Inset, general tectonic setting of the GMP region, including the approximate extent of the IR mapped before this study²¹. Red box, study area shown in **a** and **b**. Shaded bathymetric map based on our data; colour scale shows depth below sea level. Heavy black lines, axis of the EPR; thin white lines, outline of the extent of crust generated at the IR based on bathymetric lineation patterns. The lozenge shape suggests that spreading has been occurring for longest at the location of its maximum width. **b**, Crustal magnetization of the study area. Thin black lines, locations of data tracks used to calculate crustal

magnetization; data are from our cruise and from the National Geophysical Data Center (NDGC). White lines, magnetic anomaly data collected along track lines (black) to the right of each anomaly. The four western profiles were collected during our cruise; the profile to the east was obtained from NGDC. The profiles are projected such that negative anomalies plot to the west; local lows in magnetic anomaly amplitude (-150 to -200 nT) centred over the IR are consistent with recent magmatism for an east–west-oriented spreading centre at the Equator^{22,1}.

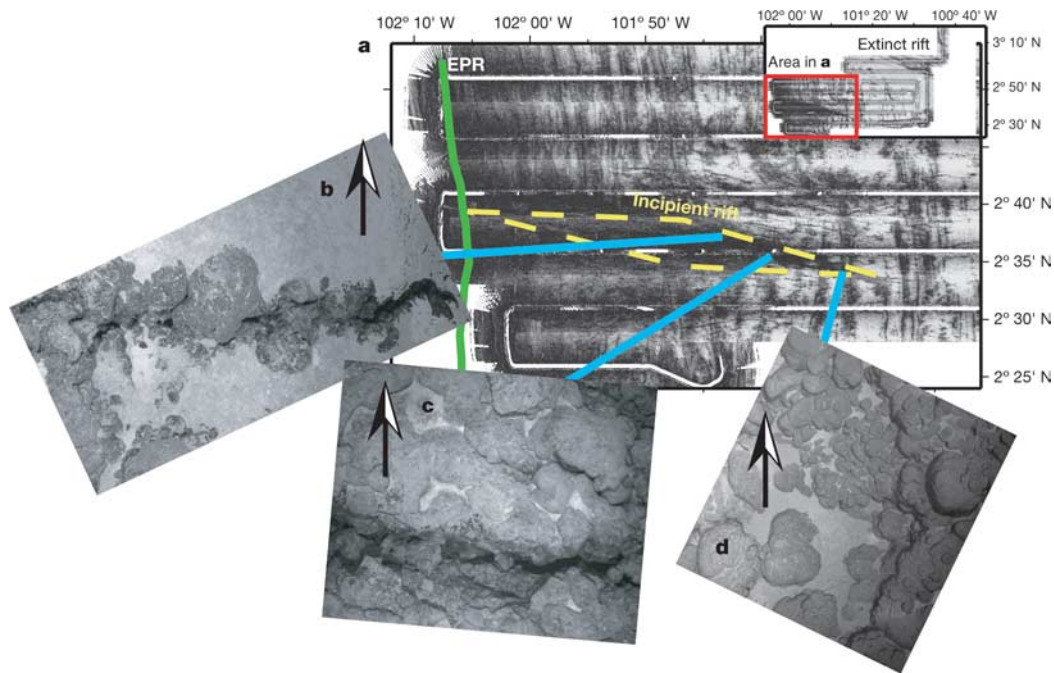


Figure 2 SeaBeam amplitude (side-scan) results and photographic images of the IR. **a**, SeaBeam amplitude data for the region in Fig. 1. Dark areas denote high-amplitude reflection, light areas show low-amplitude reflection probably due to sediment cover. Green line, EPR axis. Yellow dashed lines, outline of the extent of the crust generated at the IR from Fig. 1; the region within the lozenge shape is highly reflective compared to off-axis crust generated at the EPR. Inset, side-scan data for the broader region

including lineations of the 'extinct rift'. **b–d**, Selected photographs from camera tows along the eastern IR; arrows indicate north. Panel **b** shows an example of an east–west fissure through which lava has erupted onto older sedimented sea floor. Photographs have been rendered in black and white, and contrast and brightness enhanced for clarity.

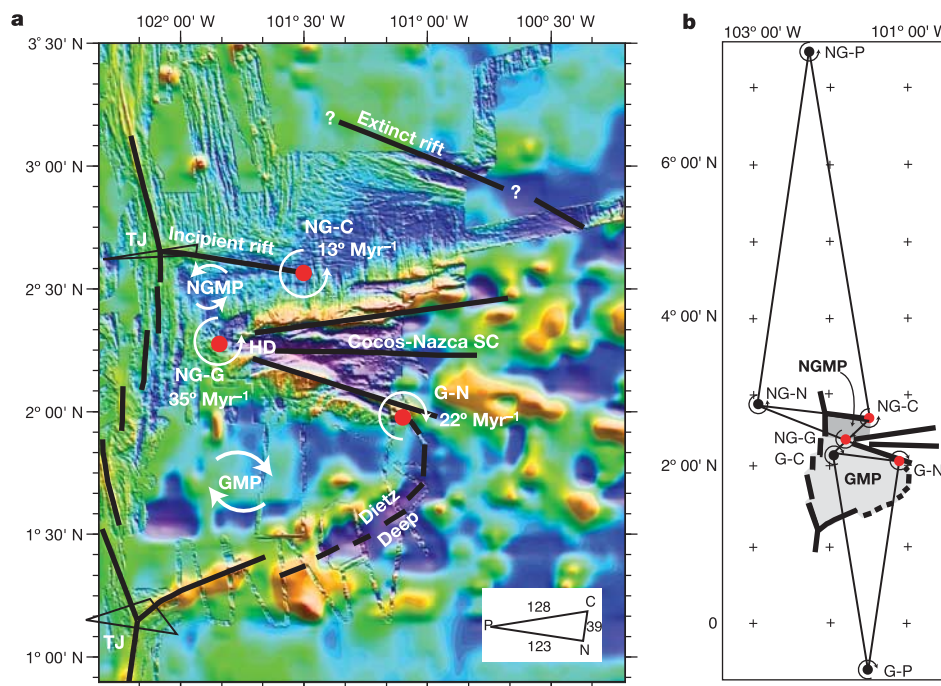


Figure 3 Tectonic configuration of the Galapagos triple junction and kinematic solution of the instantaneous motion of a dual microplate system. **a**, Bathymetric map of the Galapagos triple junction region with multibeam bathymetry on estimated seafloor topography. Black lines show spreading boundaries, Hess Deep gore, and the northwest–southeast-trending trough ('extinct rift'). Dots with circular arrows, instantaneous relative rotation axes (IRRA). IRRA at tips of rifts that bound the microplates describe motion between the NGMP and the Cocos plate (NG-C); the GMP and the Nazca plate (G-N); and

the microplates, NGMP and GMP (NG-G). Triangles, plate motion at the $1^{\circ}10'N$ and $2^{\circ}40'N$ triple junctions (TJ). Triangle in inset, relative motion of major plates¹⁹ (in km Myr^{-1}). **b**, Kinematic solution of instantaneous motion of the NGMP ($13^{\circ}\text{Myr}^{-1}$) and GMP ($22^{\circ}\text{Myr}^{-1}$) relative to Cocos, Nazca and Pacific plates. The solution consisting of 7 IRRAs satisfies the geometry of the ridge–ridge–ridge triple junctions at $1^{\circ}10'N$ and $2^{\circ}40'N$ (see Supplementary Information for details).

rotating and will be rafted away from the EPR axis. There is some evidence that this has happened in the past. A newly discovered trough, which we call the 'extinct rift', located north and east of the IR (Fig. 3a) may represent the northern boundary of such an extinct microplate.

Our new view of the Galapagos triple junction is that of two adjacent counter-rotating microplates distributing the strain around this triple junction. A kinematic solution shows that edge-driven microplate mechanisms can explain the motions of this dual microplate system. If this model is correct, we speculate that it may be applicable to other triple junctions^{17,18,20}. In the specific case of the Galapagos triple junction, we suggest that the dual microplate system acts to control the location and configuration of the Hess Deep rift and the stability of the Galapagos triple junction. □

Received 7 June; accepted 30 November 2004; doi:10.1038/nature03262.

1. Lonsdale, P. Structural pattern of the Galapagos microplate and evolution of the Galapagos triple junction. *J. Geophys. Res.* **93**, 13551–13574 (1988).
2. Lonsdale, P., Blum, N. & Puchelt, H. The RRR triple junction at the southern end of the Pacific-Cocos East Pacific Rise. *Earth Planet. Sci. Lett.* **109**, 73–85 (1992).
3. Schouten, H., Klitgord, K. D. & Gallo, D. G. Edge-driven microplate kinematics. *J. Geophys. Res.* **98**, 6689–6701 (1993).
4. Raff, A. D. Seafloor spreading: another rift. *J. Geophys. Res.* **73**, 3699–3705 (1968).
5. Hey, R. Tectonic evolution of the Cocos-Nazca spreading center. *Geol. Soc. Am. Bull.* **88**, 1404–1420 (1977).
6. Lonsdale, P. Regional shape and tectonics of the equatorial East Pacific Rise. *Mar. Geophys. Res.* **3**, 295–315 (1977).
7. Lonsdale, P. Linear volcanoes along the Pacific-Cocos plate boundary, 9°N to the Galapagos triple junction. *Tectonophysics* **116**, 255–279 (1985).
8. Searle, R. C. & Francheteau, J. Morphology and tectonics of the Galapagos triple junction. *Mar. Geophys. Res.* **8**, 95–129 (1986).
9. Baker, E. T. & Milburn, H. B. MAPR: A new instrument for hydrothermal plume mapping. *Ridge Events* **8**, 23–25 (1997).
10. Hanna, H. D., Klein, E. M., Smith, D. K. & Zhu, W. The Melville Vancouver Leg 01 Scientific Party. Along-axis geochemical variations in basaltic glasses from the Incipient Rift adjacent to the East Pacific Rise at 2°40' N. *Eos* **84**, 1495 (2003).
11. Sims, K. W. W. *et al.* Aberrant youth: chemical and isotopic constraints on the origin of off-axis lavas from the East Pacific Rise, 9°–10°N. *Geochem. Geophys. Geosyst.* **4**, doi:10.1029/2002GC000443 (2002).
12. Parker, R. L. & Huestis, S. P. The inversion of magnetic anomalies in the presence of topography. *J. Geophys. Res.* **79**, 1587–1594 (1974).
13. Macdonald, K. C., Miller, S. P., Huestis, S. P. & Spiess, F. N. Three-dimensional modeling of a magnetic reversal boundary from inversion of deep-tow measurements. *J. Geophys. Res.* **85**, 3670–3680 (1980).
14. Gee, J. & Kent, D. V. Magnetization of axial lavas from the southern East Pacific Rise (14°–23°S): Geochemical controls on magnetic properties. *J. Geophys. Res.* **102**, 24873–24886 (1997).
15. Gee, J. S., Cande, S. C., Hildebrand, J. A., Donnelly, K. & Parker, R. L. Geomagnetic intensity variations over the past 780 kyr obtained from near-seafloor magnetic anomalies. *Nature* **408**, 827–832 (2000).
16. Carey, S. W. *Continental Drift, a Symposium* 177–355 (Geology Department, University of Tasmania, Hobart, 1958).
17. Naar, D. F. & Hey, R. N. Tectonic evolution of the Easter microplate. *J. Geophys. Res.* **96**, 7961–7993 (1991).
18. Larson, R. L. *et al.* Roller-bearing tectonic evolution of the Juan Fernandez microplate. *Nature* **356**, 571–576 (1992).
19. DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Effect of recent revisions to the geomagnetic reversal timescale on estimates of current plate motions. *Geophys. Res. Lett.* **21**, 2191–2194 (1994).
20. Mitchell, N. C. Distributed extension at the Indian Ocean triple junction. *J. Geophys. Res.* **96**, 8019–8043 (1991).
21. Karson, J. A. *et al.* Structure of uppermost fast-spread oceanic crust exposed at the Hess Deep Rift: Implications for subaxial processes at the East Pacific Rise. *Geochem. Geophys. Geosyst.* **3**, doi:10.1029/2001GC001155 (2002).
22. Guyodo, Y. & Valet, J.-P. Global changes in intensity of the Earth's magnetic field during the past 800 kyr. *Nature* **399**, 249–252 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to the captain and crew of the R/V *Melville* (Vancouver Leg 01). The Incipient Rift Team included E. Klein, D. Smith, R. Cheney, R. Comer, C. Donnelly, P. Gregg, H. Hanna, G. Kurras, J. McGuire, M. Pollock, M. Rudnicki, E. Williams, C. Williams and W. Zhu. We thank G. Christeson for collecting additional bathymetry and magnetic data for us, and R. Searle and D. Wilson for comments that improved the manuscript. This work was supported by the US National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.M.K. (ek4@duke.edu).

New evidence on deinonychosaurian dinosaurs from the Late Cretaceous of Patagonia

Fernando E. Novas¹ & Diego Pol²

¹CONICET, Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", Av. Ángel Gallardo 470, Buenos Aires 1405, Argentina

²Mathematical Biosciences Institute, The Ohio State University, 231W 18th Avenue, Columbus, Ohio 43210, USA

Most of what is known about the evolution of deinonychosaurians (that is, the group of theropods most closely related to birds) is based on discoveries from North America and Asia¹. Except for *Unenlagia comahuensis*^{2,3} and some fragmentary remains from northern Africa⁴, no other evidence was available on deinonychosaurian diversity in Gondwana. Here we report a new, Late Cretaceous member of the clade, *Neuquenraptor argentinus* gen. et sp. nov., representing uncontroversial evidence of a deinonychosaurian theropod in the Southern Hemisphere. The new discovery demonstrates that Cretaceous theropod faunas from the southern continents shared greater similarity with those of the northern landmasses than previously thought. Available evidence suggests that deinonychosaurians were probably distributed worldwide at least by the beginning of the Cretaceous period. The phylogenetic position of the new deinonychosaur, as well as other Patagonian coelurosaurian theropods, is compatible with a vicariance model of diversification for some groups of Gondwanan and Laurasian dinosaurs.

Theropoda Marsh, 1881

Coelurosauria Huene, 1920

Maniraptora Gauthier, 1986

Deinonychosauria Colbert and Russell, 1969

Dromaeosauridae Matthew & Brown, 1922

Neuquenraptor argentinus gen. et sp. nov.

Etymology. *Neuquén*, a province of northwest Patagonia, and *raptor*, meaning robber in Greek; *argentinus*, in reference to Argentina.

Holotype. MCF PVPH 77 (Museo Carmen Funes, Plaza Huincul, Neuquén Province, Argentina) consists of fragments of cervical vertebra, dorsal ribs, haemal arches, left proximal radius, right femur and distal tibia, proximal tarsals, and most of the foot of the left hindlimb. It was discovered serendipitously by P. F. Puerta and F.E.N. in 1996 while digging up the rib cage of a titanosaurid sauropod.

Locality and horizon. Upper Cretaceous (Coniacian⁵), Portezuelo Formation, Sierra del Portezuelo, Neuquén Province, Argentina. Other theropods recorded in this unit are *Unenlagia*^{2,3}, *Patagonykus*⁶, *Megaraptor*⁷ and an undescribed neornithine coracoid.

Diagnosis. A probable dromaeosaurid with the following combination of characters: metatarsal II with lateral expansion over the caudal surface of metatarsal III (autapomorphic); metatarsal III proximally pinched; extensor sulcus on proximal half of metatarsus; distal end of metatarsal III is incipiently ginglymoid (to a lesser degree than other dromaeosaurids); pedal digit II with phalanges 1 and 2 sub-equal in length, and bearing a trenchant ungual phalanx.

The holotype specimen of *Neuquenraptor argentinus* (Fig. 1) is approximately 2 m long. The radius is long and gracile, with a triangular-shaped proximal articular surface, closely resembling that of *Saurornitholestes langstoni* (Museum of the Rockies, MOR 660). The femur is proportionally short and robust, similar to *Deinonychus* and *Saurornitholestes*, but different compared with

the longer and slender femur of *Unenlagia comahuensis*, thus demonstrating that MCF PVPH 77 is not a juvenile specimen of *U. comahuensis*. The fibula of *Neuquenraptor* is distally splint-like, and the calcaneum is lateromedially compressed. The preserved portion of the astragalar ascending process indicates

that it was proximodistally high, as usual among derived coelurosaurs.

Metatarsal II is transversely wider than metatarsal IV (contrasting with troodontids in which metatarsal IV is robust⁸), and ends distally in a well-developed ginglymoid articulation, a condition present in Dromaeosauridae and the basal birds *Rahonavis* and probably *Jeholornis* (X. Xu, personal communication). Metatarsal III is craniocaudally compressed, in contrast with the more robust metatarsals II and IV, which are deeper anteroposteriorly than they are mediolaterally wide. The proximal end of metatarsal III appears to be visible both cranially and caudally, as is the case with other basal deinonychosaurs. The distal end of metatarsal III expands over the cranial surfaces of metatarsals II and IV, and inversely, it is caudally hidden by lateral and medial projections of metatarsals II and IV, respectively (Fig. 1e, g). This peculiar metatarsal articulation corresponds with the arctometatarsalian condition⁹, characteristically occurring among troodontids, ornithomimids, caenagnathids, tyrannosaurids, and the dromaeosaurids *Sinornithosaurus*¹⁰ and *Microaptor*^{8,11,12}. Metatarsal IV of *Neuquenraptor* has a prominent and sharp longitudinal ridge posteriorly directed along its posterolateral margin (similar to the crista plantaris lateralis of living birds¹³), a feature absent in other paravians (that is, the clade containing deinonychosaurs and birds¹⁴), with the exception of the troodontids *Sinornithoides youngi* (Institute of Vertebrate Paleontology and Paleoanthropology, IVPP V9612) and *Sinovenator changii*^{8,15}, and the basal dromaeosaurids *Microaptor zhaoianus*^{8,11,12} and possibly *Sinornithosaurus millenii* (IVPP V12811). This is in conflict with former interpretations¹¹, according to which, in the dromaeosaurids *M. zhaoianus* and *S. millenii*, such a posterolateral ridge of metatarsal IV was interpreted as extending medially along the posteromedial edge.

An extensor sulcus exists on the cranial surface of metatarsal III, a feature also present in troodontids (for example, *Sinovenator* and *Tochisaurus*) and basal dromaeosaurids (for example, *Sinornithosaurus millenii* IVPP V12811, *Graciliraptor*⁸ and probably *Microaptor*). This extensor sulcus is absent in most theropods, including derived dromaeosaurids (for example, *Velociraptor*, *Deinonychus*, *Hulsanpes*).

Pedal digit II exhibits distinctive deinonychosaurian features (also present in the basal bird *Rahonavis*, but not in other avialans). For example, phalanges 1 and 2 have expanded distal ginglymoidal joints, and phalanx 2 bears a strong proximoventral process for extensive dorsoventral excursions. The ungual phalanx of digit II is enlarged, strongly curved, and with a sharp cutting edge. This ungual phalanx is grooved on both sides, with the lateral groove occupying a more dorsal position than the medial one—as usual among dromaeosaurids but differing from most troodontids in which the ungual phalanx is symmetrical⁸. The unguals of digits III and IV are deep, flat-bottomed, with deep collateral grooves, and with developed flexor tubercles, as in other paravians.

Character analysis (Fig. 2) depicts *Neuquenraptor* as being located basally within Deinonychosauria^{14,16}. This position is supported by the combination of an arctometatarsalian condition with a posterolateral flange on the caudal surface of metatarsal IV; a trenchant claw on the second pedal digit; and a distal end of metatarsal III that is incipiently ginglymoid. Some outstanding synapomorphies of all known troodontids (for example, asymmetrical foot with slender metatarsal II and robust metatarsal IV) or of derived troodontids (for example, tongue-like distal articular surface on metatarsal III) are absent in the new Patagonian form. In contrast, the Patagonian form exhibits a combination of pedal features (for example, metatarsal II with ginglymoidal distal end; phalanges II.1 and II.2 sub-equal in length; phalanx II.2 strongly constricted dorsoventrally at mid-shaft; and ungual of pedal digit II strongly compressed laterally and with collateral grooves asymmetrically arranged)^{4,8,17} that occur (albeit not exclusively; X. Xu, personal communication)

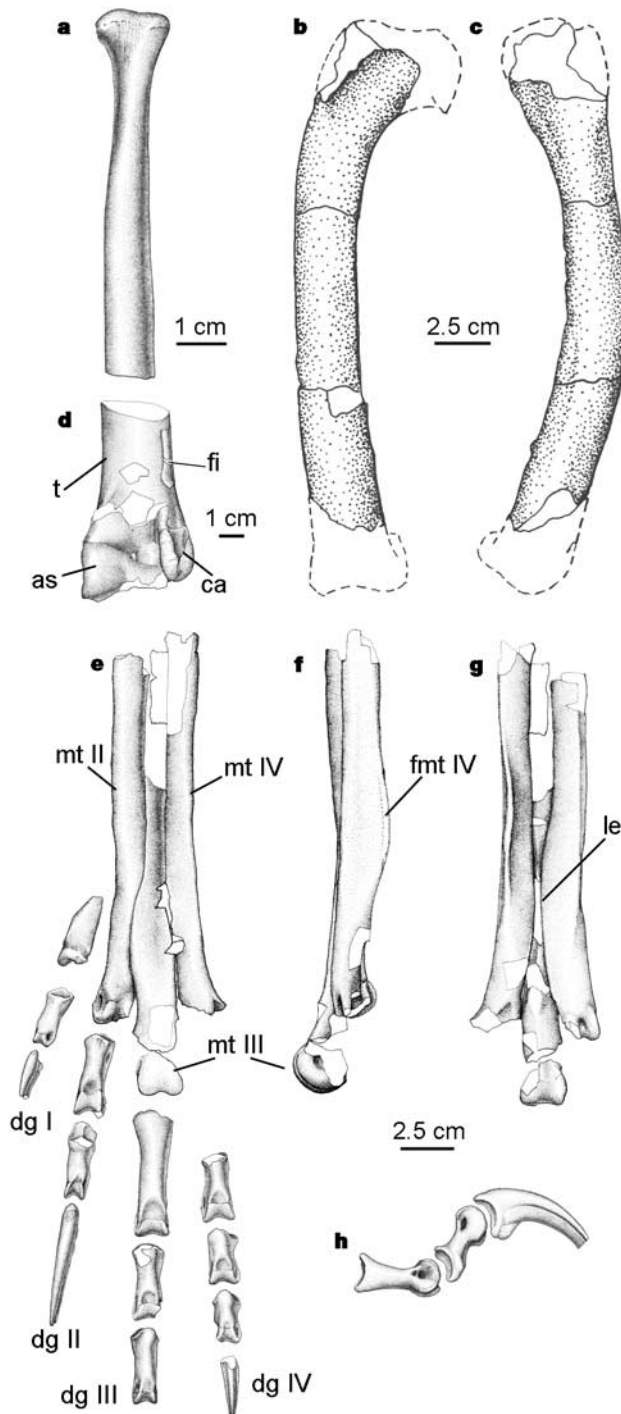


Figure 1 *Neuquenraptor argentinus*, MCF PVPH 77, holotype. **a**, Proximal end of left radius in lateral view. **b, c**, Right femur in cranial (**b**) and lateral (**c**) views. **d**, Left tibia (t), fibula (fi), astragalus (as) and calcaneum (ca) in cranial view. **e**, Left foot in cranial view. **f**, Metatarsal IV (mt IV) and III (mt III) in lateral view. **g**, Metatarsals II–IV in caudal view. **h**, Pedal digit II in medial view. dg I–IV, digits I–IV; fnt IV, posterolateral flange on metatarsal IV; le, lateral expansion of metatarsal II over caudal surface of metatarsal III.

in Dromaeosauridae, thus suggesting that *Neuquenraptor* may represent a member of this clade of deinonychosaurs (see Supplementary Information).

Neuquenraptor occupies a basal position in the Dromaeosauridae, because it lacks derived features (for example, shaft of metatarsal IV mediolaterally wide and craniocaudally flat, metatarsal III proximally broad) present in Laurasian members of this group^{1,17,18}. In other words, *Neuquenraptor* does not belong to the derived dromaeosaurid sub-clade that radiated in Laurasia, which includes *Deinonychus*, *Velociraptor*, *Dromaeosaurus*, *Adasaurus* and *Utahraptor*, among others^{1,19}, a splitting that minimally occurred during Barremian times (Fig. 2).

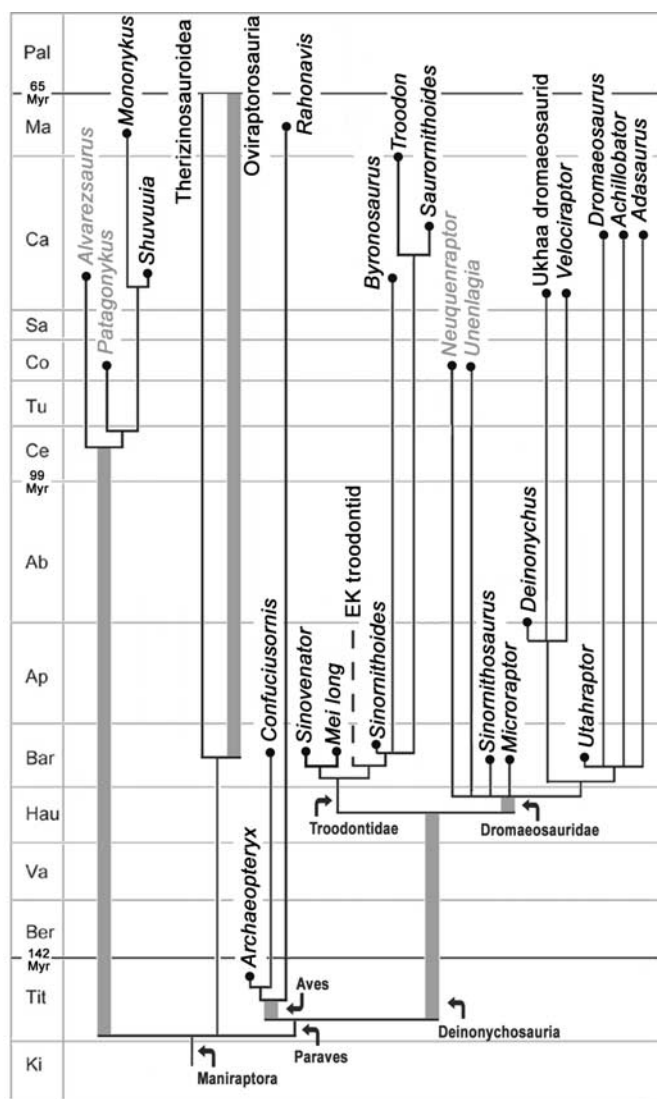


Figure 2 Reduced strict consensus depicting the phylogenetic relationships of *Neuquenraptor argentinus* within Maniraptora. Despite the large number of missing entries for *Neuquenraptor*, this taxon is unambiguously included within Dromaeosauridae based on the presence of synapomorphies of several inter-nested clades^{1,20,30}. The proposed phylogenetic tree plotted against geological time indicates that the origination of most maniraptoran lineages must have occurred no later than the Late Jurassic period. South American maniraptorans are indicated in grey font. Thick grey lines represent maniraptoran lineages with South American representatives. See Supplementary Information for further phylogenetic data. Myr, million years. Pal, Palaeocene; Ma, Maastrichtian; Ca, Campanian; Sa, Santonian; Co, Coniacian; Tu, Turonian; Ce, Cenomanian; Ab, Albian; Ap, Aptian; Bar, Barremian; Hau, Hauterivian; Va, Valanginian; Ber, Berriasian; Tit, Tithonian; Ki, Kimmeridgian.

As prompted elsewhere^{1,20}, homoplasy is a common problem in coelurosaurian phylogeny. In this regard, the arctometatarsalian metatarsus shows a complex evolutionary history, and the basal position of *Neuquenraptor* provides useful information to test the monophyly of arctometatarsalian theropods. Our analysis is consistent with recent interpretations^{1,15} that evolutionary transitions between the arctometatarsal and non-arctometatarsal foot occurred multiple times both in basal Coelurosauria (for example, Tyrannosauridae, Ornithomimidae) and maniraptorans (for example, alvarezsaurids, some oviraptorosaurs, derived troodontids and basal dromaeosaurids). The arctometatarsalian condition thus constitutes one of the homoplastic features most frequently evolved between Coelurosauria.

The discovery of *Neuquenraptor* increases the knowledge on the Gondwanan deinonychosaurs, which include the Patagonian *Unenlagia comahuensis*² and fragmentary elements from the Cenomanian of Northern Africa (Wadi Milk Formation, Sudan)⁴. The presence of these taxa in the Cretaceous period of South America and Africa dismisses the previous hypothesis²¹ claiming that deinonychosaurs were endemic from Laurasia. Moreover, the Gondwanan record of non-avian maniraptorans is also formed by presumed oviraptorosaurs²², alvarezsaurids (*Patagonykus*, *Alvarezsaurus*)^{23,24} and bizarre representatives of large size^{25,26}. This information strongly supports the hypothesis that an important adaptive radiation of maniraptoran theropods took place in the southern continents during the Cretaceous period.

Two alternative hypotheses (for example, vicariance and dispersal) may explain the presence of such a diversity of maniraptorans in Gondwana: either these lineages were descendants of Middle to Late Jurassic maniraptorans that attained a worldwide distribution, and that later produced vicariant taxa with the isolation of Gondwana from Laurasia; or that dispersal events of several maniraptoran lineages occurred later between northern and southern continents. Although dispersal of maniraptoran lineages can not be dismissed, vicariance constitutes the most parsimonious explanation for maniraptoran distribution^{4,14}. First, the presence of *Archaeopteryx* in the Tithonian demonstrates that maniraptoran diversification was well underway at the end of the Jurassic period¹. Second, the basal phylogenetic position of all known Patagonian maniraptorans (the deinonychosaurs *Neuquenraptor* and *Unenlagia*, and the alvarezsaurids *Alvarezsaurus* and *Patagonykus*) suggests a repeated cladistic pattern of area relationships compatible with a vicariance model^{27,28}. In this context, *Neuquenraptor* may represent a Late Cretaceous survivor of basal dromaeosaurids, which were probably distributed worldwide by the Late Jurassic period. Similarly, a sub-clade of derived dromaeosaurids (comprising velociraptorines plus dromaeosaurines; Fig. 2) diversified separately in Laurasia during the Cretaceous period.

The documented diversity of derived coelurosaurians shows that their evolutionary history in the southern continents is more complex than originally thought. Furthermore, this leads us to reconsider the palaeobiogeographical history of Cretaceous theropod faunas recorded in the southern landmasses, traditionally interpreted as being markedly different compared with those of the northern continents²⁹. On the basis of the new evidence, this difference still seems to apply to the large theropods (for example, abelisaurids, carcharodontosaurids, spinosaurids, tyrannosaurids), but distinctions are less important when theropods of small body size are compared (for example, deinonychosaurs, alvarezsaurids, oviraptorosaurs, basal avialans). Exceptions are gracile abelisauroid noasaurids, so far unrecorded in Laurasia. Further discoveries of Gondwanan coelurosaurians demonstrate that deinonychosaurs are among the most widely distributed clades of Cretaceous theropods. □

Received 19 July; accepted 14 December 2004; doi:10.1038/nature03285.

- Clark, J. M., Norell, M. A. & Makovicky, P. J. In *Mesozoic Birds. Above the Head of Dinosaurs* (eds Chiappe, L. M. & Witmer, L.) 31–61 (Univ. California Press, Berkeley, 2002).
- Novas, F. E. & Puerta, P. F. New evidence concerning avian origins from the Late Cretaceous of Patagonia. *Nature* **387**, 390–392 (1997).
- Novas, F. E. in *Feathered Dragons: the Origin of Birds and Flight* (eds Currie, P., Colpellhus, E. & Martin, E.) 150–166 (Indiana Univ. Press, Bloomington, 2004).
- Rauhut, O. & Werner, C. First record of the family Dromaeosauridae (Dinosauria: Theropoda) in the Cretaceous of Gondwana (Wadi Milk Formation, northern Sudan). *Palaont. Z.* **69**, 475–489 (1995).
- Leanza, H., Apesteguía, S., Novas, F. E. & de la Fuente, M. S. Cretaceous terrestrial beds from the Neuquén Basin (Argentina) and their tetrapod assemblages. *Cretaceous Res.* **25**, 61–87 (2004).
- Novas, F. E. Anatomy of *Patagonykus pueri* (Theropoda, Maniraptor), from the Late Cretaceous of Patagonia. *J. Vert. Paleontol.* **17**, 137–166 (1997).
- Novas, F. E. *Megaraptor namunhauquii*, a large-clawed, Late Cretaceous theropod from Patagonia. *J. Vert. Paleontol.* **18**, 4–9 (1998).
- Xu, X. *Deinonychosaurian Fossils from the Jehol Group of Western Liaoning and the Coelurosaurian Evolution*. Thesis, Chinese Academy of Sciences (2002).
- Holtz, T. Jr The phylogenetic position of the Tyrannosauridae: implications for theropod systematics. *J. Paleontol.* **68**, 1100–1117 (1994).
- Xu, X., Wang, X. & Wu, X. A dromaeosaurid dinosaur with a filamentous integument from the Yixian Formation of China. *Nature* **401**, 262–266 (1999).
- Xu, X., Zhou, Z. & Wang, X. The smallest known non-avian theropod dinosaur. *Nature* **408**, 705–708 (2000).
- Hwang, S. H., Norell, M. A., Ji, Q. & Gao, K. New specimens of *Microraptor zhaoianus* (Theropoda: Dromaeosauridae) from Northeastern China. *Am. Mus. Novit.* **3381**, 1–44 (2002).
- Baumel, J. J. & Witmer, L. In *Handbook of Avian Anatomy: Nomina Anatomica Avium* Vol. 23 (eds Baumel, J. J., King, A., Breazile, J., Evans, H. & Vanden Berge, J.) 45–132 (Publications Nuttall Ornithological Club, Cambridge, 1993).
- Sereno, P. C. The evolution of dinosaurs. *Science* **284**, 2137–2147 (1999).
- Xu, X., Norell, M., Wang, X.-L., Makovicky, P. J. & Wu, X. A basal troodontid from the Early Cretaceous of China. *Nature* **415**, 780–784 (2002).
- Gauthier, J. A. In *The Origin of Birds and the Evolution of Flight* (ed. Padian, K.) 1–55 (California Academy of Sciences, San Francisco, 1986).
- Norell, M. A. & Makovicky, P. J. Important features of the dromaeosaurid skeleton II: information from newly collected specimens of *Velociraptor mongoliensis*. *Am. Mus. Novit.* **3282**, 1–45 (1999).
- Hwang, S. H., Norell, M. A., Ji, Q. & Gao, K. A large compsognathid from the Early Cretaceous Yixian Formation of China. *J. System. Paleontol.* **2**, 13–30 (2004).
- Senter, P., Barsold, R., Britt, B. B. & Burnham, D. A. Systematics and evolution of Dromaeosauridae (Dinosauria, Theropoda). *Bull. Gunma Mus. Nat. Hist.* **8**, 1–20 (2004).
- Novas, F. E. & Pol, D. In *Mesozoic Birds. Above the Head of Dinosaurs* (eds Chiappe, L. M. & Witmer, L.) 121–125 (Univ. California Press, Berkeley, 2002).
- Bonaparte, J. F. Cretaceous tetrapods of Argentina. *Münchner Geowiss. Abh.* **30**, 73–130 (1996).
- Frankfurt, N. G. & Chiappe, L. M. A possible oviraptorosaur from the late Cretaceous of Northwestern Argentina. *J. Vert. Paleontol.* **19**, 101–105 (1999).
- Novas, F. E. Alvarezsauridae, Late Cretaceous maniraptorans from Patagonia and Mongolia. *Queensland Mus. Mem.* **39**, 675–702 (1996).
- Chiappe, L. M., Norell, M. A. & Clark, J. M. In *Mesozoic birds. Above the Head of Dinosaurs* (eds Chiappe, L. M. & Witmer, L.) 87–120 (Univ. California Press, Berkeley, 2002).
- Novas, F. E. & Agnolín, F. *Unquillosaurus cebalosi* Powell, a giant maniraptoran (Dinosauria, Theropoda) from the Late Cretaceous of Argentina. *Rev. Mus. Argentino Ciencias Nat.* **6**, 61–66 (2004).
- Novas, F. E., Canale, J. & Isasi, M. Giant deinonychosaurian theropod from the Late Cretaceous of Patagonia. *J. Vert. Paleontol.* **24** (suppl.), 98A (2004).
- Nelson, G. & Platnick, N. *Systematics and Biogeography: Cladistics and Vicariance* (Columbia Univ. Press, New York, 1981).
- Upchurch, P., Huxley, C. A. & Norman, D. B. An analysis of dinosaurian biogeography: evidence for the existence of vicariance and dispersal patterns caused by geological events. *Proc. R. Soc. Lond. B* **269**, 613–621 (2002).
- Bonaparte, J. F. & Kielan-Jawarowska, Z. In *Fourth Symposium on Mesozoic Terrestrial Ecosystems* (eds Currie, P. J. & Koster, E. H.) 24–29 (Tyrrell Museum of Paleontology Occasional Papers, Drumheller, 1987).
- Xu, X. & Norell, M. A new troodontid dinosaur from China with avian-like sleeping posture. *Nature* **431**, 838–841 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank X. Xing, O. Rauhut, M. Norell and P. Makovicky for comments and discussion on this subject; R. A. Coria for the loan of *Neuquenraptor argentinus* specimens; M. Norell for access to new specimens of *Velociraptor mongoliensis*; J. Ostrom and J. A. Gauthier for access to *Deinonychus antirrhopus*; X. Xing and P. Currie for access to several maniraptoran specimens; A. Scanferla for technical preparation of the specimen; and J. González for the illustrations. Fieldwork was supported by the National Geographic Society. This study was sponsored by Conicet, Agencia Nacional de Promoción Científica y Tecnológica, The Dinosaur Society, The Jurassic Foundation, Akapol SA, and Renault Argentina (Buenos Aires).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.E.N. (fernovas@yahoo.com.ar).

Prokaryotic cells of the deep sub-seafloor biosphere identified as living bacteria

Axel Schippers¹, Lev N. Neretin^{1,2}, Jens Kallmeyer^{2,3}, Timothy G. Ferdelman², Barry A. Cragg⁴, R. John Parkes⁴ & Bo B. Jørgensen²

¹Section Geomicrobiology, Federal Institute for Geosciences and Natural Resources, Stilleweg 2, 30655 Hannover, Germany

²Department of Biogeochemistry, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany

³GeoForschungsZentrum Potsdam, PB 4.3, Telegrafenberg, 14473 Potsdam, Germany

⁴School of Earth, Ocean and Planetary Sciences, Cardiff University, Main Building, Park Place, Cardiff CF10 3YE, Wales, UK

Chemical analyses of the pore waters from hundreds of deep ocean sediment cores have over decades provided evidence for ongoing processes that require biological catalysis by prokaryotes^{1–3}. This sub-seafloor activity of microorganisms may influence the surface Earth by changing the chemistry of the ocean and by triggering the emission of methane, with consequences for the marine carbon cycle and even the global climate^{4–6}. Despite the fact that only about 1% of the total marine primary production of organic carbon is available for deep-sea microorganisms^{7,8}, sub-seafloor sediments harbour over half of all prokaryotic cells on Earth⁷. This estimation has been calculated from numerous microscopic cell counts in sediment cores of the Ocean Drilling Program^{1,9}. Because these counts cannot differentiate between dead and alive cells, the population size of living microorganisms is unknown^{10,11}. Here, using ribosomal RNA as a target for the technique known as catalysed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH), we provide direct quantification of live cells as defined by the presence of ribosomes. We show that a large fraction of the sub-seafloor prokaryotes is alive, even in very old (16 million yr) and deep (>400 m) sediments. All detectable living cells belong to the Bacteria and have turnover times of 0.25–22 yr, comparable to surface sediments.

Direct evidence for the existence of a deep biosphere is provided by the following: (1) microscopic cell counts using unspecific fluorescent DNA (RNA) stains such as acridine orange; (2) sequences of high-molecular-weight prokaryotic DNA; (3) cultivation of diverse bacteria from subsurface sediments; and (4) bacterial activities measured with radiotracers¹. Only a minute fraction of the enumerated cells were so far culturable and it has therefore remained unknown what fraction of cells is alive and active. The fluorochrome acridine orange, routinely applied in microscopic cell counting, binds unspecifically to DNA and RNA and thus does not provide information on the viability of the cells^{10,11}. Potentially, a large part of the counted cells could be dormant or even dead and yet retain stainable DNA. RNA, in contrast, is much more labile and is readily degraded in cells that become inactive due to starvation. Cell death in pure cultures accelerates when less than half of the RNA remains¹¹. Starved cells may still maintain an intact cell membrane and nucleic acids such as DNA or transfer RNA, but they rapidly lose their ribosomes¹². The experience from pure culture studies is that cells with a significant ribosome content are living and metabolically active. We therefore used a highly sensitive molecular technique targeting specifically rRNA as an indicator of living cells in deeply buried marine sediments. The technique CARD-FISH was combined with quantitative, real-time polymerase chain reaction (Q-PCR) quantification of 16S ribosomal DNA genes, to determine what fraction of prokaryotic cells

that had been quantified by acridine orange direct counting (AODC) in the subsurface was indeed alive according to the above criteria.

Representative for the open ocean and the ocean margin, deep subsurface sediments with different geochemical regimes were explored in the Ocean Drilling Program (ODP) Leg 201 (refs 2, 13). The sampling sites comprised a variety of Miocene to Holocene sediments in the open ocean (sites 1225 and 1226) and in the ocean margin (sites 1227 and 1230) of the eastern tropical Pacific Ocean. The oldest sediments immediately overlie basaltic basement and have a biostratigraphic age of 11 Myr for site 1225 and 16.5 Myr for site 1226. Carbonate and siliceous oozes were cored at these sites at water depths of 3,760 m and 3,297 m, respectively. Biogenic oozes and terrigenous sediments of the shallow Peru shelf were cored at site 1227 at a water depth of 427 m. Organic-rich sediments containing gas hydrate were cored in the accretionary wedge at site 1230 in the Peru Trench at 5,086 m water depth. In the ocean-margin sediments the average concentration of total organic carbon (TOC) was in the range of 1–10% dry weight (dw). The TOC content was one to two orders of magnitude lower in the open-ocean sediments. The sediment temperature varied between 1 and 26 °C, the range for psychrophilic and mesophilic micro-organisms¹³.

The Q-PCR data for the ocean-margin sites exhibited almost identical numbers of total prokaryotes and Bacteria, and one to three times fewer Archaea (Fig. 1). The finding shows that Bacteria are the dominant prokaryotes in the Leg 201 ocean-margin sediments. The numbers of total prokaryotes and Bacteria decreased from 10^8 cells cm^{-3} at the top of the core to about 10^6 cells cm^{-3} at 40 m below the seafloor (mbsf) for both sites. Below this depth, the abundance of Bacteria and of total prokaryotes fluctuated around 10^5 cells cm^{-3} for site 1227 and 10^6 cells cm^{-3} for site 1230. The decrease of Archaea within the top 40 mbsf was more pronounced than that of Bacteria. Whereas at the top up to 10^7 cells cm^{-3} of Archaea were found, the values decreased to less than 10^4 cells cm^{-3} at 40 mbsf for both sites.

Bacteria were detected at the open-ocean and ocean-margin sites using CARD-FISH (Fig. 2). Depth profiles of AODC and numbers of Bacteria determined by CARD-FISH and Q-PCR showed that AODC counts were generally higher at the ocean-margin sites than at the open-ocean sites, consistent with earlier results^{1,9} (Fig. 3). A large proportion of the AODC counts was detected by CARD-FISH analysis: about one-third for the open-ocean and up to one-tenth

for the ocean-margin sediments. All depth profiles of bacterial numbers determined by CARD-FISH did not show a significant decrease in cell numbers with depth, in contrast to AODC. Our data are the first demonstration of high numbers of living Bacteria in deeply buried marine sediments. The CARD-FISH numbers represent a minimum of the total living Bacteria in the deep sediments because living cells with very low activity and ribosome contents were probably not detected. Because we do not have any information about the activity and the ribosome content of single cells in the deeply buried sediments, our definition of CARD-FISH-targeted Bacteria as living Bacteria is a utilitarian definition. At all sites the abundance of Archaea was too low to be quantified using CARD-FISH, confirming together with our Q-PCR data that

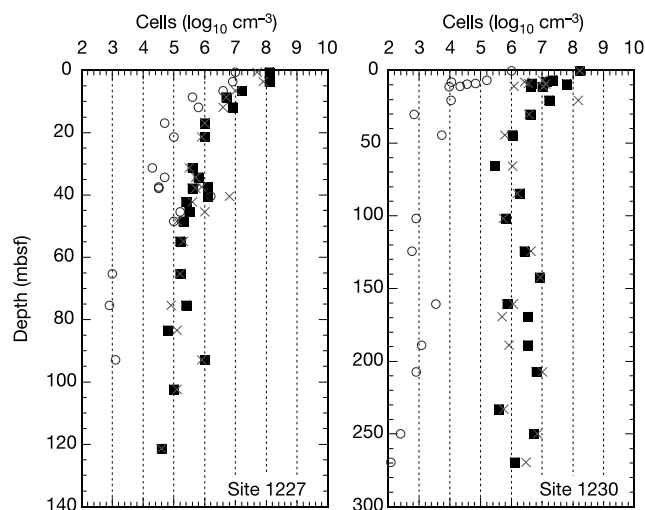


Figure 1 Depth profiles of total prokaryotes (squares), Bacteria (crosses) and Archaea (circles) determined by Q-PCR for two ocean-margin sites.

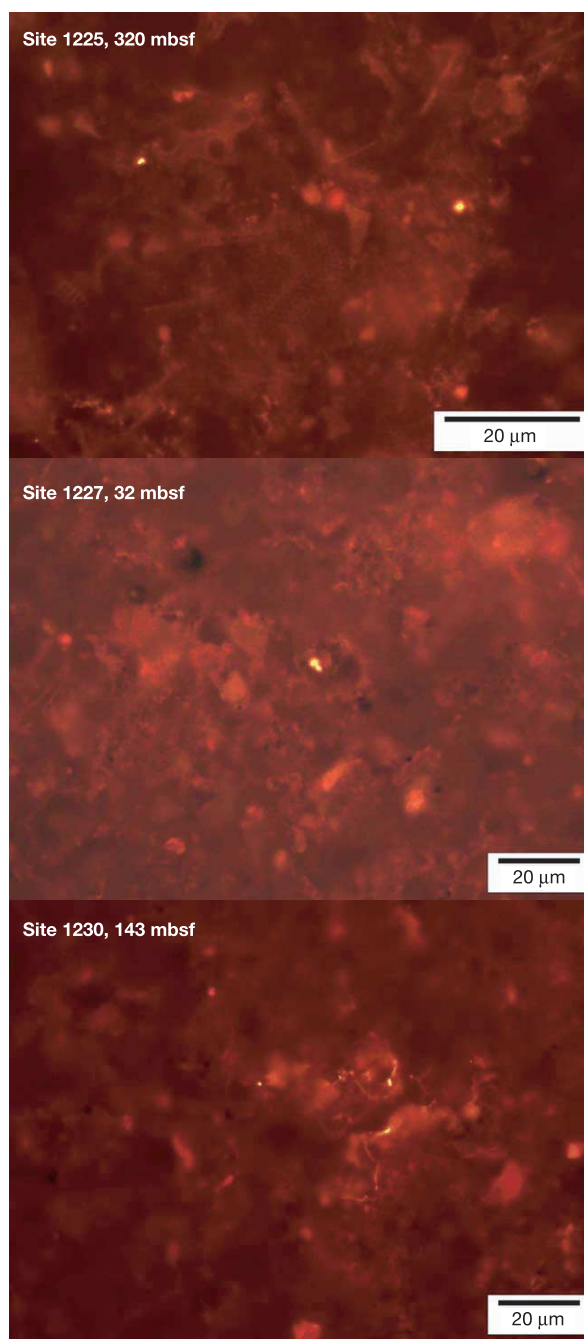


Figure 2 Bacteria detected by CARD-FISH for one open-ocean (site 1225) and two ocean-margin sites (sites 1227 and 1230).

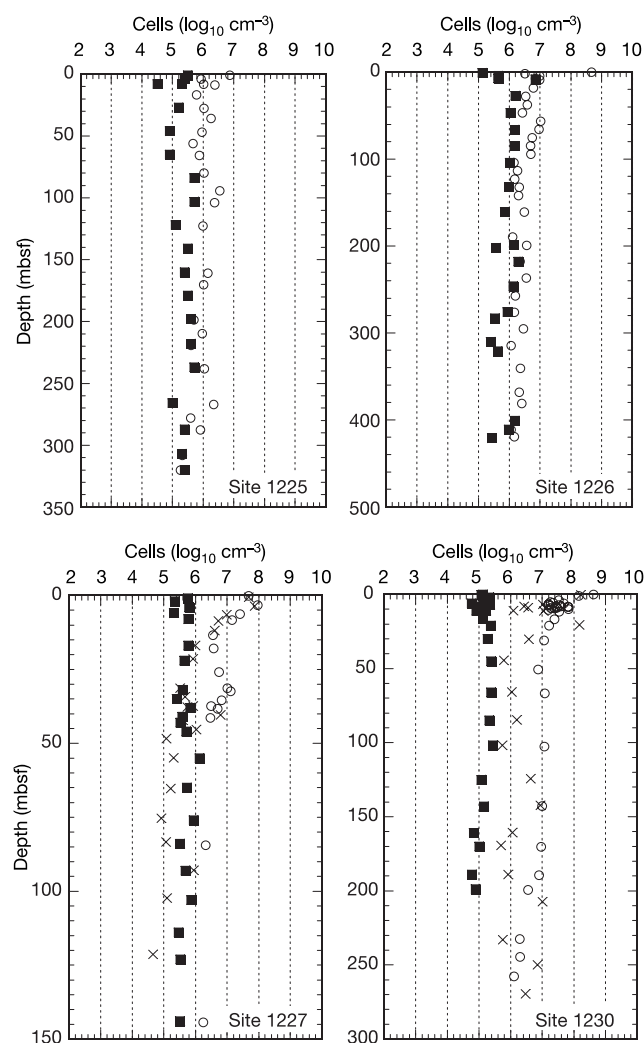


Figure 3 Depth profiles of AODC and numbers of Bacteria. AODC (circles) and numbers of Bacteria were determined by CARD-FISH (squares) and Q-PCR (crosses) for two open-ocean and two ocean-margin sites. For each site, the mean standard deviation of the CARD-FISH counts was 0.4 cells ($\log_{10} \text{cm}^{-3}$).

Bacteria are the dominant prokaryotes in deeply buried marine sediments.

The continuous decrease of AODC with depth together with the almost constant depth distribution of living Bacteria over the entire sediment column reflects the ongoing degradation of cells with sediment depth and age. The Q-PCR bacterial numbers decrease with depth and age as well, showing the degradation of high-molecular-weight DNA targeted by our Q-PCR method. The

Q-PCR cell numbers were obtained indirectly assuming that the average 16S rDNA copy number per cell is 3.6 (ref. 14). The difference between AODC and the number of Bacteria determined by Q-PCR may be biased by variable 16S rRNA operon numbers and/or variable genome copy numbers for different bacterial taxa^{14,15}.

The difference between CARD-FISH-detectable, living Bacteria and total cell numbers (AODC) is much higher at the ocean-margin than at the open-ocean sites. This difference may be explained by the greater availability of different electron acceptors in the open-ocean sediments, where sulphate, Fe(III), Mn(IV) and even nitrate are available as electron acceptors for microbial respiration. By contrast, only sulphate is present down to about 40 and 10 mbsf for the ocean-margin sites 1227 and 1230, respectively^{2,13}. Below these depths, methanogenesis by Archaea is the main terminal pathway of organic carbon mineralization. An alternative explanation would be better preservation of DNA in the ocean margin-sediments, where organic carbon is also better preserved.

Organic matter degradation by fermentation and anaerobic respiration are the principal energy delivering processes in marine subsurface sediments^{2,13}. According to the depth profiles of sulphate, Fe(II), Mn(II), and nitrate in interstitial water, sulphate reduction is the most important terminal mineralization process for the four sites of this study^{2,13}, as has been previously described for surface sediments¹⁶.

We measured gross sulphate reduction rates (SRR) using ³⁵S-radiotracer and modelled net SRRs based on sulphate depth profiles. The potential biomass formation by sulphate reduction and the numbers of living Bacteria were used to calculate turnover times of bacterial biomass in the subsurface (Table 1). The areal SRRs were higher for the ocean-margin than for the open-ocean sites and, as expected, the measured gross SRR was mostly higher than the modelled net SRR. The turnover times of bacteria were in the range of 0.25–1.91 yr, both for the open-ocean and for the ocean-margin sites. Higher turnover times for living bacterial biomass of 7 yr for ocean-margin and 22 yr for open-ocean sediments were calculated from the global estimates of carbon flux available for the subsurface bacterial community and the total living bacterial biomass. All these values are comparable to turnover times of prokaryotes in soil and aquatic habitats and are considerably lower than the value of $1-2 \times 10^3$ yr given by ref. 7 for the turnover time of the total prokaryotic biomass in subsurface sediments.

From the number of living Bacteria, we calculated the total number and biomass of living prokaryotes in the oceanic subsurface, which equal 1.3×10^{29} cells and 2.5×10^{15} g of cellular C with an almost identical contribution of open-ocean and ocean-margin sediments. According to the CARD-FISH data, the number of living prokaryotes on Earth is, by an order of magnitude, fewer than the number of total prokaryotes previously calculated on the basis of AODC⁷. Because not all living Bacteria may be detected by CARD-FISH, this is a minimum estimate. This study shows that subsurface marine sediments, as one of the least active environments on Earth,

Table 1 Potential C oxidation by sulphate reduction and calculation of turnover times

Pacific Ocean area	ODP Leg 201 site	Relevant depth interval for sulphate reduction (m)	Potential C oxidation by modelled sulphate reduction* ($\text{mmol m}^{-2} \text{yr}^{-1}$)	Potential C oxidation by measured sulphate reduction† ($\text{mmol m}^{-2} \text{yr}^{-1}$)	Potential biomass formation by measured sulphate reduction‡ ($\text{mg cellular C m}^{-2} \text{yr}^{-1}$)	Turnover time§ for living Bacteria (yr)
Open ocean	1225	1–317	0.29	ND	ND	ND
Open ocean	1226	1–17	3.5	27	67.5	1.91
Ocean margin	1227	1–34	52	161	403	0.82
Ocean margin	1230	1–11	50	40	100	0.25

*Rates calculated using the model and method of Berg *et al.*²⁵

†Rates derived from shipboard ³⁵S tracer experiments.

‡Calculation based on the free energy value of $-100 \text{ kJ mol C}^{-1}$ and the Gibbs energy dissipation coefficient of 40 kJ g^{-1} cellular C^{29,30}.

§Calculation based on median cell numbers for the relevant depth interval and using the value of $19 \text{ fg C per cell}^{11}$. ND, not determined.

contain a high number of living Bacteria with turnover rates comparable to surface environments. □

Methods

Sampling

Samples for AODC and CARD-FISH analyses were taken as subsamples from the centremost part of the core immediately after the core was brought on deck. AODC counts were partly determined onboard and partly determined after the cruise¹³, and samples for CARD-FISH were fixed onboard for analysis after the cruise. For Q-PCR analysis, 5-cm-long sections were frozen at -80°C . Potential contamination with seawater microorganisms was routinely checked by application of fluorescent beads that were the size of prokaryotic cells and of perfluorocarbon as tracers during drilling¹³. Only uncontaminated samples were used for analyses.

CARD-FISH

Samples for CARD-FISH analysis were fixed immediately after sampling¹⁷ and stored in ethanol: PBS (1:1) at -20°C for post-cruise analyses. CARD-FISH was applied on filters following a previously described protocol for marine bacteria¹⁸. For each sample, the filter was cut into sections that were used for hybridization, targeting either Archaea (probe ARCH915), Bacteria (probe EUB338), or no cells (probe NON338 as negative control)¹⁷. The standard deviation of three individual counts per probe was calculated for each sample and a mean standard deviation was calculated for each site (Fig. 3). The efficiency and reproducibility of our CARD-FISH method was confirmed by an almost complete recovery of bacterial cells previously added in different numbers to samples from all ODP sites. Spores (as shown for spores of *Bacillus licheniformis*) could not be visualized, most probably because their cell walls were not sufficiently permeabilized with the CARD-FISH protocol¹⁸. In contrast, vegetative cells of the same strain were detected.

Additionally, as control experiments three further CARD-FISH applications were performed for five to ten selected samples each: (1) *Planctomycetales* are not targeted by the probe EUB338 (ref. 17). Instead we used the probe PLA886 (ref. 17) to quantify *Planctomycetales*, but cells could not be detected in contrast to the positive control organism *Pirellula* sp.; (2) an improved protocol for quantification of gram-positive Bacteria¹⁹ was applied. Bacterial numbers of the ODP samples were not statistically different to those obtained with the previous protocol¹⁸; and (3) an improved protocol for quantification of Archaea²⁰ was applied. The abundance of Archaea in the ODP samples was too low to be quantified, as has been found with the previous protocol¹⁸.

Q-PCR

For Q-PCR analysis, high-molecular-weight DNA was extracted from 1–5 g of each frozen sediment sample from the centremost part following a modified FastDNA Spin Kit for Soil (Bio101) protocol²¹. Q-PCR (ABI Prism 7000 or 7700, Applied Biosystems) was used to determine the 16S rDNA copy numbers of total prokaryotes²², Archaea²² and Bacteria²³. 16S rDNA gene copy numbers were converted to cell numbers using a conversion factor of 3.6 (ref. 14). The DNA recovery efficiency and reproducibility of our Q-PCR method was confirmed by an almost complete recovery of bacterial cells previously added in different numbers to samples from all ODP sites.

Sulphate reduction rates

Experimental determination of bacterial sulphate reduction rates (SRR) were performed onboard on triplicate syringe sub-core samples. Briefly, the sub-cores were injected with 10 μl of a ^{35}S -sulphate tracer-containing solution (50 $\text{kBq } \mu\text{l}^{-1}$) and incubated for 14 to 33 days. Details of the onboard incubation and onshore processing in Bremen are given elsewhere^{13,24}. Rates and detection limits were calculated as previously described²⁴. Areal organic carbon turnover rates were estimated by summing SRR over the depth range relevant to sub-surface sulphate reduction and using a S:C stoichiometry of 1:2. Modelled rates of sulphate reduction were numerically obtained using a statistical curve-fitting approach developed for biogeochemical interpretation²⁵ and these rates closely match those flux estimates derived by other approaches⁷.

Total number of active Bacteria and turnover time based on carbon flux

The total number of living Bacteria in the oceanic subsurface sediments was determined from median CARD-FISH numbers per site from this study and from the volumes of subsurface sediments given by ref. 7 separately for open-ocean and ocean-margin sites. For sediments deeper than 400 m, the same number of living Bacteria was used as above, taking into account the constancy of CARD-FISH profiles over depth at our sites. The total amount of carbon stored in Bacteria in the subsurface was estimated from their numbers and a cellular carbon content of 19 fg C cell⁻¹. This cellular carbon content was obtained by summarizing available literature data for cellular carbon, DNA and protein contents in aquatic and soil bacteria^{11,26–28}. The resulting values were 7.3×10^{28} cells and 1.4×10^{15} g C for open-ocean, and 5.6×10^{28} cells and 1.1×10^{15} g C for ocean-margin subsurface sediments.

The turnover time of living Bacteria was calculated by dividing the carbon flux available for the subsurface community by the total number of living Bacteria, estimated as described above, separately for the open-ocean and ocean-margin sites. We assumed that 1% of the total primary production in both (4×10^{14} g C yr⁻¹ in the open-ocean and 1×10^{14} g C yr⁻¹ in the ocean-margin sites) minus the C burial rate (5×10^{12} g C yr⁻¹ and 10×10^{12} g C yr⁻¹ for open-ocean and ocean-margins, respectively⁹) is available for subsurface microorganisms⁷. A carbon assimilation efficiency of 0.5 (ref. 29) was used to calculate the turnover times.

Received 30 July; accepted 16 December 2004; doi:10.1038/nature03302.

1. Parkes, R. J., Cragg, B. A. & Wellsbury, P. Recent studies on bacterial populations and processes in subseafloor sediments: a review. *Hydrogeol. J.* **8**, 11–28 (2000).
2. D'Hondt, S. *et al.* Distributions of microbial activities in deep subseafloor sediments. *Science* **306**, 2216–2221 (2004).
3. Holland, H. D. *The Chemical Evolution of the Atmosphere and Oceans* (Princeton Univ. Press, Princeton, 1984).
4. Dickens, G. R., O'Neil, J. R., Rea, D. K. & Owen, R. M. Dissociation of oceanic methane hydrate as a cause of the carbon isotope excursion at the end of the Paleocene. *Paleoceanography* **10**, 965–971 (1995).
5. Kennett, J. P., Cannariato, K. G., Hendy, I. L. & Behl, R. J. Carbon isotopic evidence for methane hydrate instability during Quaternary interstadials. *Science* **288**, 128–133 (2000).
6. Hesselbo, S. P. *et al.* Massive dissociation of gas hydrate during a Jurassic oceanic anoxic event. *Nature* **406**, 392–395 (2000).
7. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
8. Hedges, J. I. Global biogeochemical cycles: progress and problems. *Mar. Chem.* **39**, 67–93 (1992).
9. Parkes, R. J. *et al.* Deep bacterial biosphere in Pacific Ocean sediments. *Nature* **371**, 410–413 (1994).
10. Kepner, R. L. Jr & Pratt, J. R. Use of fluorochromes for direct enumeration of total bacteria in environmental samples: past and present. *Microbiol. Rev.* **58**, 603–615 (1994).
11. Morita, R. Y. (ed.) *Bacteria in Oligotrophic Environments* (Chapman & Hall, New York, 1997).
12. Davis, B. D., Luger, S. M. & Tai, P. C. Role of ribosome degradation in the death of starved *Escherichia coli* cells. *J. Bacteriol.* **166**, 439–445 (1986).
13. D'Hondt, S. L. *et al.* Controls on Microbial Communities in Deeply Buried Sediments, Eastern Equatorial Pacific and Peru Margin (Ocean Drilling Program, http://www-odp.tamu.edu/publications/201_IR/201ir.htm, 2003).
14. Klappenbach, J. L., Saxman, P. R., Cole, J. R. & Schmidt, T. M. rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res.* **29**, 181–184 (2001).
15. Fogel, G. B., Collins, C. R., Li, J. & Brunk, C. F. Prokaryotic genome size and SSU rDNA copy number: estimation of microbial relative abundance from a mixed population. *Microb. Ecol.* **38**, 93–113 (1999).
16. Jørgensen, B. B. Mineralization of organic matter in the sea bed—the role of sulphate reduction. *Nature* **296**, 643–645 (1982).
17. Ravenschlag, K., Sahm, K. & Amann, R. Quantitative molecular analysis of the microbial community in marine Arctic sediments. *Appl. Environ. Microbiol.* **67**, 387–395 (2001).
18. Pernthaler, A., Pernthaler, J. & Amann, R. Fluorescence *in situ* hybridization and catalyzed reporter deposition for the identification of marine bacteria. *Appl. Environ. Microbiol.* **68**, 3094–3101 (2002).
19. Sekar, R. *et al.* An improved protocol for quantification of freshwater actinobacteria by fluorescence *in situ* hybridization. *Appl. Environ. Microbiol.* **69**, 2928–2935 (2003).
20. Teira, E., Reinthaler, T., Pernthaler, A., Pernthaler, J. & Herndl, G. J. Combining catalyzed reporter deposition-fluorescence *in situ* hybridization and microautoradiography to detect substrate utilization by Bacteria and Archaea in the deep ocean. *Appl. Environ. Microbiol.* **70**, 4411–4414 (2004).
21. Webster, G., Newberry, C. J., Fry, J. C. & Weightman, A. J. Assessment of bacterial community structure in the deep sub-seafloor biosphere by 16S rDNA-based techniques: a cautionary tale. *J. Microbiol. Methods* **55**, 155–164 (2003).
22. Takai, K. & Horikoshi, K. Rapid detection and quantification of members of the archaeal community by quantitative PCR using fluorogenic probes. *Appl. Environ. Microbiol.* **66**, 5066–5072 (2000).
23. Nadkarni, M., Martin, F. E., Jacques, N. A. & Hunter, N. Determination of bacterial load by real-time PCR using a broad range (universal) probe and primer set. *Microbiol.* **148**, 257–266 (2002).
24. Kallmeyer, J., Ferdelman, T. G., Weber, A., Fossing, H. & Jørgensen, B. B. A cold chromium distillation procedure for radiolabeled sulfide applied to sulfate reduction measurements. *Limnol. Oceanogr. Methods* **2**, 171–180 (2004).
25. Berg, P., Risgaard-Petersen, N. & Rysgaard, S. Interpretation of measured concentration profiles in sediment pore water. *Limnol. Oceanogr.* **43**, 1500–1510 (1998).
26. Løferer-Krössbacher, M., Witzel, K.-P. & Psenner, P. DNA content of aquatic bacteria measured by densitometric image analysis. *Arch. Hydrobiol. Spec. Issues Adv. Limnol.* **54**, 185–198 (1999).
27. Simon, M. & Azam, F. Protein content and protein synthesis rates of planktonic marine bacteria. *Mar. Ecol. Prog. Ser.* **51**, 201–213 (1989).
28. Zubkov, M. V., Fuchs, B. M., Burkil, P. H. & Amann, R. Comparison of cellular and biomass specific activities of dominant bacterioplankton groups in stratified waters of the Celtic Sea. *Appl. Environ. Microbiol.* **67**, 5210–5218 (2001).
29. Heijnen, J. J. & van Dieken, J. P. In search of a thermodynamic description of biomass yields for the chemotrophic growth of microorganisms. *Biotechnol. Bioeng.* **39**, 833–858 (1992).
30. Bach, W. & Edwards, K. J. Iron and sulfide oxidation within the basaltic ocean crust: implications for chemolithoautotrophic microbial biomass production. *Geochim. Cosmochim. Acta* **67**, 3871–3887 (2003).

Acknowledgements We thank the ODP Leg 201 personnel and shipboard scientists for sampling and discussions, especially F. Inagaki and A. Teske. This research was supported by a grant to A.S., T.G.F. and B.B.J. from the priority program IODP/ODP of the German Research Foundation (DFG).

Authors' contributions A.S. and L.N.N. formulated the main ideas as a result of discussions with T.G.F. and B.B.J., and had the main responsibility for writing the Letter. A.S. did CARD-FISH and Q-PCR analysis, the latter together with L.N.N. B.A.C. and R.J.P. provided AODC data, and J.K., T.G.F. and B.B.J. provided sulphate reduction rates.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.S. (a.schippers@bgr.de).

The genome of the protist parasite *Entamoeba histolytica*

Brendan Loftus¹, Iain Anderson¹, Rob Davies², U. Cecilia M. Alsmark³, John Samuelson⁴, Paolo Amedeo¹, Paola Roncaglia¹, Matt Berriman², Robert P. Hirt³, Barbara J. Mann⁵, Tomo Nozaki⁶, Bernard Suh¹, Mihai Pop¹, Michael Duchene⁷, John Ackers⁸, Egbert Tannich⁹, Matthias Leippe¹⁰, Margit Hofer⁷, Iris Bruchhaus⁹, Ute Willhoeft⁹, Alok Bhattacharya¹¹, Tracey Chillingworth², Carol Churcher², Zahra Hance², Barbara Harris², David Harris², Kay Jagels², Sharon Moule², Karen Mungall², Doug Ormond², Rob Squares², Sally Whitehead², Michael A. Quail², Ester Rabinowitsch², Halina Norbertczak², Claire Price², Zheng Wang¹, Nancy Guillén¹², Carol Gilchrist³, Suzanne E. Stroup³, Sudha Bhattacharya¹¹, Anuradha Lohia¹³, Peter G. Foster¹⁴, Thomas Sicheritz-Ponten¹⁵, Christian Weber¹², Upinder Singh¹⁶, Chandrama Mukherjee¹³, Najib M. El-Sayed¹, William A. Petri Jr³, C. Graham Clark⁸, T. Martin Embley³, Bart Barrell², Claire M. Fraser¹ & Neil Hall^{2*}

¹TIGR, 9712 Medical Center Drive, Rockville, Maryland 20850, USA

²The Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

³School of Biology, University of Newcastle, King George VI Building, Newcastle upon Tyne NE1 7RU, UK

⁴Department of Molecular and Cell Biology, Boston University Goldman School of Dental Medicine, 715 Albany Street, Boston, Massachusetts 02118, USA

⁵Departments of Internal Medicine & Microbiology, University of Virginia, Charlottesville, Virginia 22908, USA

⁶Department of Parasitology, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan

⁷Division of Specific Prophylaxis and Tropical Medicine, Center for Physiology and Pathophysiology, Medical University of Vienna, Kinderspitalgasse 15, A-1095 Vienna, Austria

⁸Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

⁹Department of Molecular Parasitology, Bernhard Nocht Institute for Tropical Medicine, Bernhard Nocht Str. 74, 20359 Hamburg, Germany

¹⁰Zoological Institute, University of Kiel, Olshausenstr. 40, 24098 Kiel, Germany

¹¹School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110067, India

¹²Unité de Biologie Cellulaire du Parasitisme, INSERM U389, Institut Pasteur 28, rue du Dr Roux 75724, Paris Cedex 15, France

¹³Department of Biochemistry, Bose Institute, P1/12 CIT Scheme VIIM, Kolkata 700054, India

¹⁴Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

¹⁵Center for Biological Sequence Analysis, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

¹⁶Departments of Internal Medicine, Microbiology, and Immunology, Stanford University School of Medicine, Stanford, California 94305-5107, USA

* Present address: TIGR, 9712 Medical Center Drive, Rockville, Maryland 20850, USA

Entamoeba histolytica is an intestinal parasite and the causative agent of amoebiasis, which is a significant source of morbidity and mortality in developing countries¹. Here we present the genome of *E. histolytica*, which reveals a variety of metabolic adaptations shared with two other amitochondrial protist pathogens: *Giardia lamblia* and *Trichomonas vaginalis*. These adaptations include reduction or elimination of most mitochondrial metabolic pathways and the use of oxidative stress enzymes generally associated with anaerobic prokaryotes. Phylogenomic analysis identifies evidence for lateral gene transfer of bacterial genes into the *E. histolytica* genome, the effects of which centre on expanding aspects of *E. histolytica*'s metabolic repertoire. The presence of these genes and the potential for novel metabolic pathways in *E. histolytica* may allow for the development of new chemotherapeutic agents. The genome encodes a large number of novel receptor kinases and contains expansions of a variety of

gene families, including those associated with virulence. Additional genome features include an abundance of tandemly repeated transfer-RNA-containing arrays, which may have a structural function in the genome. Analysis of the genome provides new insights into the workings and genome evolution of a major human pathogen.

Genome analysis was carried out on a 12.5-fold coverage genome assembly consisting of 23,751,783 base pairs (bp) distributed among 888 scaffolds. The 9,938 predicted genes average 1.17 kilobases (kb) in size and comprise 49% of the genome. One-quarter of *E. histolytica* genes are predicted to contain introns, with 6% of genes containing multiple introns. No homologues could be identified for a third of predicted proteins (31.8%) from the public databases (see Methods). *E. histolytica* chromosomes do not condense, and the uncertainty surrounding its ploidy and the extensive length variability observed between homologous chromosomes from different isolates makes the exact chromosome number difficult to determine. The chromosome size variation observed may be due to expansion and contraction of subtelomeric repeats, as in other protists^{2,3}, and it is tempting to speculate that in *E. histolytica* these regions consist of tRNA-containing arrays. Comprising almost 10% of the sequence reads, 25 types of long tandem array, each containing between one and five tRNA types per repeat unit, could be identified from the genome data. The full complement of tRNAs required for translation has been identified, and all but four of the tRNA genes are encoded exclusively in arrays. These unique tRNA gene arrays are thus predicted to be functional as well as potentially fulfilling a structural role in the genome. No association could be determined between codon usage and the relative copy numbers of their cognate tRNA species.

The metabolism of *E. histolytica* seems to have been shaped by secondary gene loss and lateral gene transfer (LGT), primarily from bacterial lineages (Fig. 1). *E. histolytica* is an obligate fermenter, using bacterial-like fermentation enzymes and lacking proteins of the tricarboxylic acid cycle and mitochondrial electron transport chain. An atrophic, mitochondrion-derived organelle has been identified in *E. histolytica*⁴, and the genome data support the absence of a mitochondrial genome. Glucose is the main energy source; however, in place of the typical eukaryotic glucose transporters those of *E. histolytica* are related to the prokaryote glucose/ribose porter family, with the amino- and carboxy-terminal domains switched relative to their prokaryotic counterparts.

As a phagocytic resident of the human gut, *E. histolytica* has access to many bacterial and host-derived preformed organic compounds. Most pathways for amino acid biosynthesis have been eliminated, except those for serine and cysteine, which are probably retained for the production of cysteine, the major intracellular thiol. The high levels of cysteine in *E. histolytica* may compensate for the lack of glutathione and its associated enzymes, a major component of oxidative stress resistance in many organisms⁵. *E. histolytica* lacks *de novo* purine, pyrimidine and thymidylate synthesis and must rely on salvage pathways, similar to *G. lamblia* and *T. vaginalis*⁶. In addition, *E. histolytica* appears to lack ribonucleotide reductase, a characteristic that it shares with *G. lamblia*⁷. *E. histolytica* is unable to synthesize fatty acids but retains the ability to synthesize a variety of phospholipids. The absence of identifiable pathways for the synthesis of isoprenoids and the sphingolipid head group aminoethylphosphonate suggest the existence of novel pathways. These pathways, once characterized, might represent attractive drug targets. Two unusual enzymes of fatty acid elongation are shared between *E. histolytica* and *G. lamblia*, including a predicted acetyl-CoA carboxylase with two carboxyltransferase domains⁸. We propose that this enzyme removes a carboxyl group from oxaloacetate and transfers it to acetyl-CoA to form malonyl-CoA and pyruvate. *E. histolytica* also has five members of a fatty acid elongase family, previously identified only in plants, green algae and *G. lamblia*^{9,10}. Folate is

a cofactor essential for thymidylate synthesis and methionine recycling, and genome analysis reveals a complete lack of genes coding for known folate-dependent enzymes and folate transporters. Folate is also required for organelle protein synthesis in mitochondria and chloroplasts, and loss of the mitochondrial genome may have paved the way for the loss of these folate-dependent functions.

LGT is an important force in the evolution of prokaryotes but significantly less is known about its importance in eukaryotic evolution¹¹. We conducted a phylogenetic screen of the *Entamoeba* genome for cases of relatively recent prokaryote to eukaryote LGT (see Methods), and for 96 genes we believe that this is the simplest explanation for the tree topologies obtained (see Supplementary Information). These genes are embedded among typically eukaryotic genes on *E. histolytica* scaffolds and do not seem to represent contaminating prokaryotic sequences. Most (58%) of the LGT genes encode a variety of metabolic enzymes, whereas most of the remaining genes (41%) encode proteins of unknown function (Supplementary Fig. 1). The major impact is in the area of carbohydrate and amino acid metabolism, where they have increased the range of substrates available for energy generation including tryptophanase and aspartase, which contribute to the use of amino acids. Several glycosidases and sugar kinases appear to have been acquired through LGT and would probably enable *E. histolytica* to use sugars other than glucose; for example, fructose and galactose. There is a strong bias in the data for a major donor being in the *Cytophaga-Flavobacterium-Bacteroides* (CFB) group of the phylum Bacteroidetes; however, this should be interpreted with caution, as current sampling of prokaryotic genomes is still relatively incomplete. It is clear that among the 96 genes, some result in significant enhancements to *E. histolytica* metabolism, thus contributing to its biology to a greater extent than indicated by the numbers alone.

E. histolytica feed on bacteria in the lumen of the colon and lyse host epithelial cells after invasion of the intestinal wall¹². A number of amoebic virulence determinants have been characterized, including a multi-subunit GalGalNAc lectin involved in adhesion to host cells, cysteine proteases that degrade host extracellular matrix, and pore-forming peptides (amoebapores) capable of lysing target cells¹². Analysis of the genome reveals redundancy in the genes encoding these virulence factors. Thirty homologues of the intermediate subunit and one homologue of the heavy subunit of the GalGalNAc lectin were identified. Ten new cysteine proteinases with predicted N-terminal transmembrane anchors, which might allow them to be localized on the amoeba cell surface, were identified. In addition to three new amoebapores a homologue of haemolysin III was identified, suggesting that, in addition to amoebapores, haemolysins may have a role in host cell lysis.

Vesicle trafficking has a role in *E. histolytica* pathogenesis through phagocytosis and the delivery of secreted hydrolytic enzymes and amoebapores to the cell surface¹³. *E. histolytica* lacks morphologically identifiable rough endoplasmic reticulum and the Golgi apparatus¹⁴ but encodes the basic elements of the vesicle transport machinery common to other eukaryotic cells, with the coat complexes COPI, COPII, clathrin and retromer all being present. Rab and Arf protein family expansions reflect the increased complexity and number of vesicle fusion and recycling steps that have been associated with phagocytosis and pinocytosis in amoebae¹⁵. The cytoskeleton has a number of important roles in parasite motility, contact-dependant killing and phagocytosis of host intestinal epithelial cells¹⁶. This is reflected in expansions of Rho GTPases and their regulators RhoGAPs and RhoGEFs, which control a number of processes involving the actin cytoskeleton. Five proteins with a unique domain architecture containing both RhoGEF and Arf-GAP domains were identified, suggesting a mechanism for direct

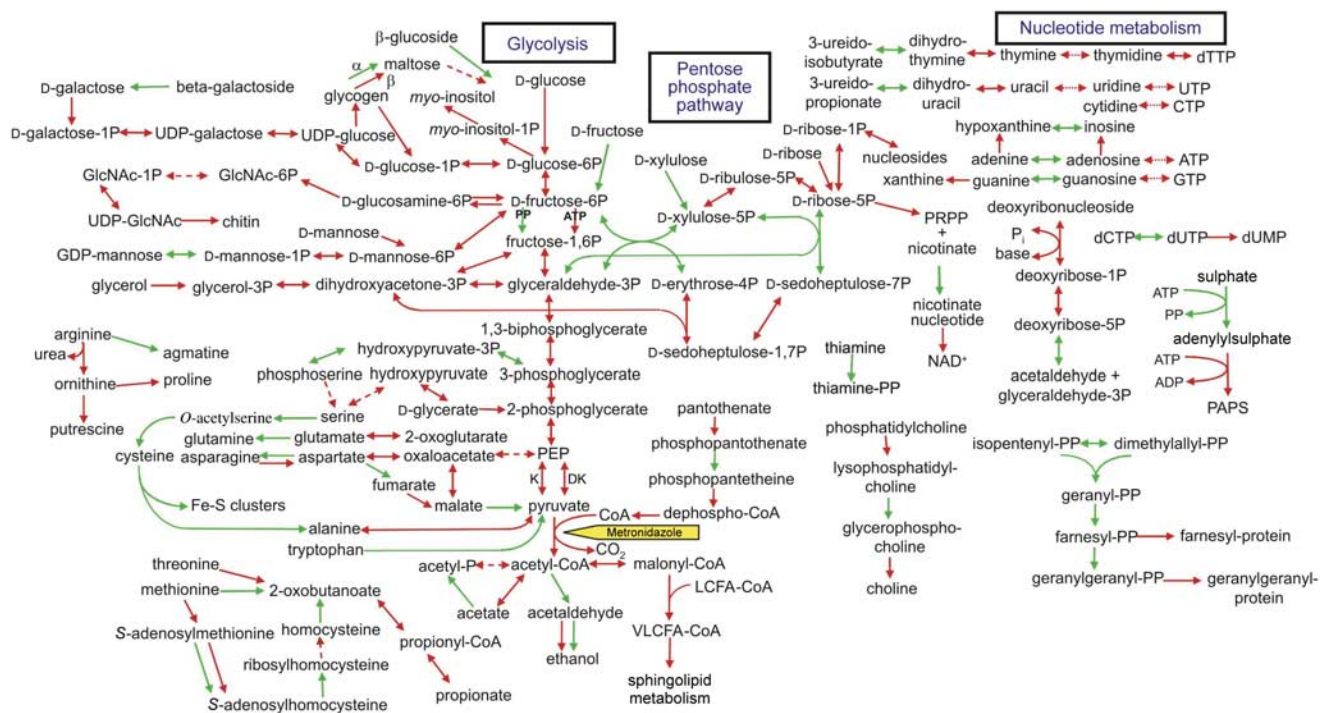


Figure 1 Predicted metabolism of *E. histolytica* based on analysis of the genome sequence data. Arrows indicate enzyme reactions. Glycolysis and fermentation are the major energy generation pathways. Green arrows represent enzymes encoded by genes that are among the 96 candidates for LGT into the *E. histolytica* genome. Broken arrows indicate enzymes for which no gene could be identified using searches of the genome data, although the activity is likely to be present. The yellow arrow points to the source of

electrons for activation of metronidazole, the major drug for treatment of amoebic liver abscess. DK, pyruvate phosphate dikinase; GlcNAc, *N*-acetylglucosamine; GPI, glycosylphosphatidylinositol; K, pyruvate kinase; LCFA, long-chain fatty acid; PAPS, phosphoadenosine phosphosulphate; PEP, phosphoenolpyruvate; PP, pyrophosphate; PRPP, phosphoribosyl pyrophosphate; VLCFA, very-long-chain fatty acid.

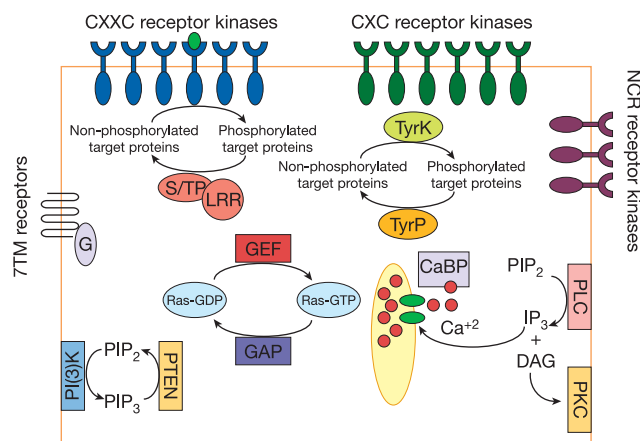


Figure 2 Predicted signal transduction mechanisms of *E. histolytica* based on analysis of the genome sequence data. *E. histolytica* possesses three types of receptor serine/threonine kinases: one group has CXXC repeats in the extracellular domain; a second has CXC repeats; and a third has non-cysteine rich (NCR) repeats. *E. histolytica* has cytosolic tyrosine kinases (TyrK), but not receptor tyrosine kinases. Some serine/threonine phosphatases (S/TP) have an attached LRR domain. CaBP, calcium-binding protein; DAG, diacylglycerol; G, G protein; GAP, GTPase-activating protein; GEF, guanine nucleotide exchange factor; IP₃, inositol-1,4,5-trisphosphate; PI(3)K, phosphatidylinositol-3-OH kinase; PIP₂, phosphatidylinositol-4,5-bisphosphate; PIP₃, phosphatidylinositol-3,4,5-trisphosphate; PKC, protein kinase C; PLC, phospholipase C; PTEN, phosphatase and tensin homologue; TyrP, tyrosine phosphatase; TTM receptors, seven-transmembrane receptors.

communication between the regulators of vesicle budding and cytoskeletal rearrangement.

E. histolytica uses a complex mix of signal transduction systems in order to sense and interact with the different environments it encounters (Fig. 2). Almost 270 putative *E. histolytica* protein kinases representing members of all seven families of the eukaryotic protein kinase superfamily were identified¹⁷. These include tyrosine kinases with SH2 domains, tyrosine kinase-like protein kinases and 90 putative receptor Ser/Thr kinases. These Ser/Thr kinases are uncommon in protists, appear to be absent from *Dictyostelium* and have previously been described only in plants, animals and Choanoflagellates. The *E. histolytica* receptor Ser/Thr kinases all contain an N-terminal signal peptide, a predicted extracellular domain and a single transmembrane helix followed by a cytosolic tyrosine kinase-like domain. The receptor kinases fall into three groups on the basis of differences in their predicted extracellular domains. The first group of 50 receptor kinase proteins contains CXXC-rich repeats similar to those found in the intermediate subunit (Igl) of the Gal/GalNAc lectin and *G. lamblia* variant-specific surface proteins. A second group of 32 proteins encodes cysteine-rich domains containing CXC repeats. The third group of eight receptor kinase-like proteins lacks cysteine-rich extracellular domains. Although no immediate downstream effectors to the amoebic receptor kinases could be identified, *E. histolytica* contains greater than 100 protein phosphatases, which dephosphorylate proteins. An unusual feature of some of the phosphatases is the presence of varying numbers of leucine-rich repeat (LRR) domains that are involved primarily in protein-protein interactions and have not previously been associated with phosphatases. The *E. histolytica* genome encodes numerous putative seven-transmembrane receptors and trimeric G proteins, which are probably involved in mediating autocrine stimulation of encystation¹⁸. In contrast to autocrine stimulation of *Dictyostelium* sporulation, which uses secreted cyclic AMP, *E. histolytica* encystation is self-stimulated by secreted catecholamines¹⁸. Finally, *E. histolytica* has numerous cytosolic proteins involved in signal transduction, including Ras-family proteins,

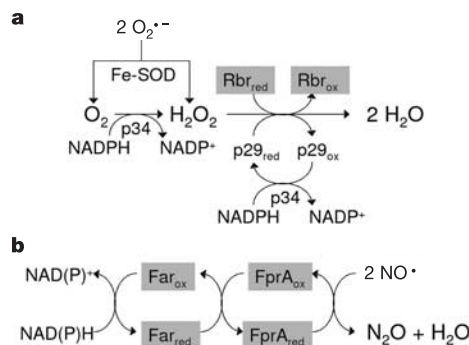


Figure 3 Predicted pathways for oxidative and nitrosative stress resistance in *E. histolytica*. Enzymes boxed and shaded have previously only been identified in anaerobic prokaryotes and amitochondrial protists. **a**, Superoxide is detoxified by an iron-containing superoxide dismutase (Fe-SOD). Molecular oxygen is reduced to hydrogen peroxide by the NADPH-flavin oxidoreductase (p34), which also transfers electrons to peroxiredoxin (p29). Rubrerythrin (Rbr) is predicted to convert hydrogen peroxide to water, although the source of electrons for rubrerythrin in *E. histolytica* is unknown. **b**, A-type flavoproteins (FprA) detoxify nitric oxide to nitrous oxide. FprA receives electrons from flavoprotein A reductase (Far).

EF-hand calcium-binding proteins, phosphatidylinositol-3-OH kinase and MAP kinases. This represents the most varied set of signal-transduction-related proteins yet described in a single-celled eukaryote.

In contrast to life in the anoxic colon, *E. histolytica* encounters a relatively high-oxygen environment during invasive amoebiasis, and coping with this change is therefore an important virulence factor. The importance of this response is underscored by the redundancy of oxygen detoxification mechanisms. *E. histolytica* has four copies of flavoprotein A, which detoxifies nitric oxide and/or oxygen¹⁹ (Fig. 3), and also contains rubrerythrin, which in anaerobic bacteria is protective against intracellular hydrogen peroxide²⁰ (Fig. 3). These oxidative and/or nitrosative stress resistance genes are shared with *G. lamblia* (with the exception of rubrerythrin) and *T. vaginalis*, but have generally been associated with anaerobic prokaryotes (Fig. 3).

E. histolytica is the first amoeba genome to be fully sampled, and comparisons with other genomes will assist in resolving fundamental issues relating to eukaryote and amoeba phylogeny, as well as how LGT affects eukaryotes. Despite a lack of representative genome sampling from amitochondrial protist lineages it is already clear that these unrelated anaerobic eukaryotes seem to use convergent metabolic strategies imposed by their environments. As a first insight into an amitochondrial protist genome, analysis of these data and particularly the bacterial-like proteins contained therein should illuminate future efforts aimed at the development of diagnostics and therapeutics of these luminal parasites. □

Methods

Genome sequencing and assembly

The *E. histolytica* genome sequence was generated by the whole-genome shotgun method. As the chromosomes of *E. histolytica* could not be resolved by pulsed field gel electrophoresis (PFGE) and the A + T content precluded making large or medium insert libraries in bacterial artificial chromosomes (BACs), we were required to use the whole-genome shotgun approach to sequence the genome. Genomic DNA was prepared from *E. histolytica* strain HM-1:IMSS (ATCC number 30459) grown axenically in TYI-S-33 medium²⁰. At TIGR 390,000 reads were produced from a small (1.5–2.0 kb) and a medium insert library (8–10 kb) generated in the pHOS2 vector. At the Sanger Institute, 200,000 reads were generated from a pUC18 library with average insert size of 2.5 kb plus 6,500 reads from a BAC library with an average insert size of 10 kb (the high A + T content of the genomic DNA prevented cloning of larger fragments). To avoid assembly problems, reads containing episomal-derived rDNA or tRNA-containing sequences (170,000 reads (29%)) were excluded from the whole-genome assembly process. The average edited read length was 645 bp, giving an approximate 12.5-fold genome coverage. Genome assembly was carried out at the Sanger Centre using the program phusion²¹. All scaffolds smaller than

2 kb (327) were subsequently removed, leaving 1,425 scaffolds with a combined size of 25,393,225 bp. The remaining scaffolds were analysed to remove redundancy that may have resulted as a consequence of allelic differences or aneuploidy. We removed all scaffolds smaller than 5 kb that shared 98% or more nucleotide sequence identity over greater than 95% of their lengths. Removal of these scaffolds left 888 scaffolds remaining, with a total length of 23,751,783 bp. All scaffolds removed during the clean-up process as well as any singleton reads, although not used in the annotation process, were used in determining the presence or absence of genes in the *E. histolytica* genome. Unfortunately, there is no map to order the scaffolds generated by the assembly; however, the sequence generated by this project should assist in making maps for this genome in the future, and although the large-scale structure of the genome has been lost, the vast majority of the genes that have been predicted are full length with intact 3' and 5' untranslated regions.

Annotation

The Combiner algorithm was used for gene structure identification²² using two genefinder programs, phat²³ and GlimmerHMM²⁴, trained using a set of published *E. histolytica* gene sequences, alignments of protein homologues to the genomic sequence and alignment of a set of *E. histolytica* complementary DNA sequences (provided by N. Guillén) to the genomic sequence. The Combiner gene predictions were then manually curated. Functional annotations for the predicted proteins were automatically generated using a combination of numerous sources of evidence including searches against a non-redundant protein database and identification of functional domains by searches against the Pfam database²⁵. tRNAs were detected using the tRNAscan-SE²⁶ program with default parameters.

Identification of sequence homologues in other species

Sequence homologues from other species were identified by searching the predicted proteins from the *E. histolytica* genome against the publicly available nr database of GenBank using BlastP (<http://www.ncbi.nlm.nih.gov/BLAST/>) and filtering search results with an *e*-value of 10^{-5} or less, which was chosen because of the relatively large divergence between *E. histolytica* and other organisms for which the genomes have been sequenced and for which protein data are available.

Phylogenetic analysis

We modified a published suite of scripts and modules called PyPhy²⁷ to make an automated genome-wide primary screen for LGT. PyPhy was used to make bootstrap (100 replicates) consensus *p*-distance trees from edited alignments of 5,740 *E. histolytica* proteins; that is, those for which there were sufficient homologues (>4) in SwissProt and TrEMBL to make trees. The trees were analysed to identify cases where the nearest neighbour to the *E. histolytica* protein was a prokaryotic sequence. As an additional screen for LGT we identified all proteins for which a prokaryote was the top Blast hit. After manual inspection of the alignments, Blast outputs, tree support values and sequence identities, 279 cases of potential LGT were retained for more detailed phylogenetic analyses. Each candidate LGT was analysed by MrBayes²⁸ using the WAG matrix, a gamma correction for site rate variation and a proportion (pinvar) of invariant sites. The analyses were run for 600,000 generations and sampled every 100 generations, with the first 2,000 samples discarded as burn-in. A consensus tree was made from the remaining samples. Because posterior probabilities—the support values used by bayesian analysis to indicate confidence in groups—have been criticized²⁹, we also used bootstrapping to provide an additional indication of support for relationships. Each data set was bootstrapped (100 replicates) and used to make distance matrices under the same evolutionary model as in the bayesian analysis, using custom (P4) software (available on request). Trees were made from the distance matrices using FastME³⁰ and a bootstrap consensus tree made using P4. On the basis of these analyses we identified 96 genes in which the tree topology is consistent with prokaryote to eukaryote LGT. Blast summary statistics, trees and support values for these 96 candidate LGT are provided as Supplementary Information.

Received 26 October; accepted 2 December 2004; doi:10.1038/nature03291.

1. Stanley, S. L. Jr Amoebiasis. *Lancet* **361**, 1025–1034 (2003).
2. Patarapotikul, J. & Langley, G. Chromosome size polymorphism in *Plasmodium falciparum* can involve deletions of the subtelomeric pFrep20 sequence. *Nucleic Acids Res.* **16**, 4331–4340 (1988).
3. Melville, S. E., Gerrard, C. S. & Blackwell, J. M. Multiple causes of size variation in the diploid megabase chromosomes of African trypanosomes. *Chromosome Res.* **7**, 191–203 (1999).
4. Leon-Avila, G. & Tovar, J. Mitosomes of *Entamoeba histolytica* are abundant mitochondrion-related remnant organelles that lack a detectable organellar genome. *Microbiology* **150**, 1245–1250 (2004).
5. Fahey, R. C., Newton, G. L., Arrick, B., Overdank-Bogart, T. & Aley, S. B. *Entamoeba histolytica*: a eukaryote without glutathione metabolism. *Science* **224**, 70–72 (1984).
6. Abrahamsen, M. S. *et al.* Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304**, 441–445 (2004).
7. Baum, K. F., Berens, R. L., Marr, J. J., Harrington, J. A. & Spector, T. Purine deoxynucleoside salvage in *Giardia lamblia*. *J. Biol. Chem.* **264**, 21087–21090 (1989).
8. Jordan, I. K., Henze, K., Fedorova, N. D., Koonin, E. V. & Galperin, M. Y. Phylogenomic analysis of the *Giardia intestinalis* transcarboxylase reveals multiple instances of domain fusion and fission in the evolution of biotin-dependent enzymes. *J. Mol. Microbiol. Biotechnol.* **5**, 172–189 (2003).
9. James, D. W. Jr *et al.* Directed tagging of the *Arabidopsis* fatty acid elongation1 (FAE1) gene with the maize transposon activator. *Plant Cell* **7**, 309–319 (1995).
10. Azachi, M. *et al.* Salt induction of fatty acid elongase and membrane lipid modifications in the extreme halotolerant alga *Dunaliella salina*. *Plant Physiol.* **129**, 1320–1329 (2002).
11. Lawrence, J. G. & Hendrickson, H. Lateral gene transfer: when will adolescence end? *Mol. Microbiol.* **50**, 739–749 (2003).
12. Huston, C. D. Parasite and host contributions to the pathogenesis of amebic colitis. *Trends Parasitol.* **20**, 23–26 (2004).
13. Welter, B. H. & Temesvári, L. A. A unique Rab GTPase, EhRabA, of *Entamoeba histolytica*, localizes to the leading edge of motile cells. *Mol. Biochem. Parasitol.* **135**, 185–195 (2004).

14. Mazzucco, A., Benchimol, M. & De Souza, W. Endoplasmic reticulum and Golgi-like elements in *Entamoeba*. *Micron* **28**, 241–247 (1997).
15. Duhon, D. & Cardelli, J. The regulation of phagosome maturation in *Dictyostelium*. *J. Muscle Res. Cell Motil.* **23**, 803–808 (2002).
16. Voigt, H. & Guillen, N. New insights into the role of the cytoskeleton in phagocytosis of *Entamoeba histolytica*. *Cell. Microbiol.* **1**, 195–203 (1999).
17. Hunter, T. Protein kinase classification. *Methods Enzymol.* **200**, 3–37 (1991).
18. Coppi, A., Merali, S. & Eichinger, D. The enteric parasite *Entamoeba* uses an autocrine catecholamine system during differentiation into the infectious cyst stage. *J. Biol. Chem.* **277**, 8083–8090 (2002).
19. Gomes, C. M. *et al.* A novel type of nitric-oxide reductase. *Escherichia coli* flavoreductoxin. *J. Biol. Chem.* **277**, 25273–25276 (2002).
20. Sztukowska, M., Bugno, M., Potempa, J., Travis, J. & Kurtz, D. M. Jr Role of rubrerythrin in the oxidative stress response of *Porphyromonas gingivalis*. *Mol. Microbiol.* **44**, 479–488 (2002).
21. Mullikin, J. C. & Ning, Z. The phusion assembler. *Genome Res.* **13**, 81–90 (2003).
22. Allen, J. E., Pertea, M. & Salzberg, S. L. Computational gene prediction using multiple sources of evidence. *Genome Res.* **14**, 142–148 (2004).
23. Cawley, S. E., Wirth, A. I. & Speed, T. P. Phat—a gene finding program for *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **118**, 167–174 (2001).
24. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
25. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
26. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
27. Sicheritz-Ponten, T. & Andersson, S. G. A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* **29**, 545–552 (2001).
28. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
29. Cummings, M. P. *et al.* Comparing bootstrap and posterior probability values in the four-taxon case. *Syst. Biol.* **52**, 477–487 (2003).
30. Desper, R. & Gascuel, O. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol. Biol. Evol.* **21**, 587–598 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by grants from the National Institute of Allergy and Infectious Disease and the Wellcome Trust.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.L. (bjloftus@tigr.org). Scaffold sequences have been deposited in GenBank under the project accession number AAFB00000000. Sequences and annotation are available at <http://www.tigr.org/tdb/e2k1/eha1/>.

Excitatory cortical neurons form fine-scale functional networks

Yumiko Yoshimura*, Jami L. M. Dantzker* & Edward M. Callaway

Systems Neurobiology Laboratories, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA

* Present addresses: Department of Visual Neuroscience, Research Institute of Environmental Medicine, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan (Y.Y.); Department of Neurology and Neurological Sciences, Stanford University, 300 Pasteur Drive, Room M016, Stanford, California 94305-5122, USA (J.L.M.D.)

The specificity of cortical neuron connections creates columns of functionally similar neurons spanning from the pia to the white matter^{1–6}. Here we investigate whether there is an additional, finer level of specificity that creates subnetworks of excitatory neurons within functional columns. We tested for fine-scale specificity of connections to cortical layer 2/3 pyramidal neurons in rat visual cortex by using cross-correlation analyses of synaptic currents evoked by photostimulation. Recording simultaneously from adjacent layer 2/3 pyramidal cells, we find that when they are connected to each other (20% of all recorded pairs) they share common input from layer 4 and within layer 2/3. When adjacent layer 2/3 neurons are not connected to each other, they share very little (if any) common excitatory input from layers 4 and 2/3. In contrast, all layer 2/3 neurons share common excitatory input

from layer 5 and inhibitory input from layers 2/3 and 4, regardless of whether they are connected to each other. Thus, excitatory connections from layer 4 to layer 2/3 and within layer 2/3 form fine-scale assemblies of selectively interconnected neurons; inhibitory connections and excitatory connections from layer 5 link neurons across these fine-scale subnetworks. Relatively independent subnetworks of excitatory neurons are therefore embedded within the larger-scale functional architecture; this allows neighbouring neurons to convey information more independently than suggested by previous descriptions of cortical circuitry.

The cerebral cortex consists of a complex network of neuronal connections, the organization of which is believed to contribute

critically to perception and behaviour. Over the last several decades, the idea of the 'functional column' has provided a dominant influence on studies of the organization and function of cortical circuits¹. The specificities of connections that both create and maintain functional architecture are well established¹⁻⁷ and provide a substrate for interactions between neurons with similar functional attributes. However, each excitatory neuron connects to only a minority of others in the same column^{8,9}. This sparse connectivity is consistent with two different scenarios, each of which has implications for the way that cortical circuits process information. In the first scenario, connections would be dependent on the spatial overlap of dendrites and axons, but would otherwise be determined probabilistically, independent of other connections in

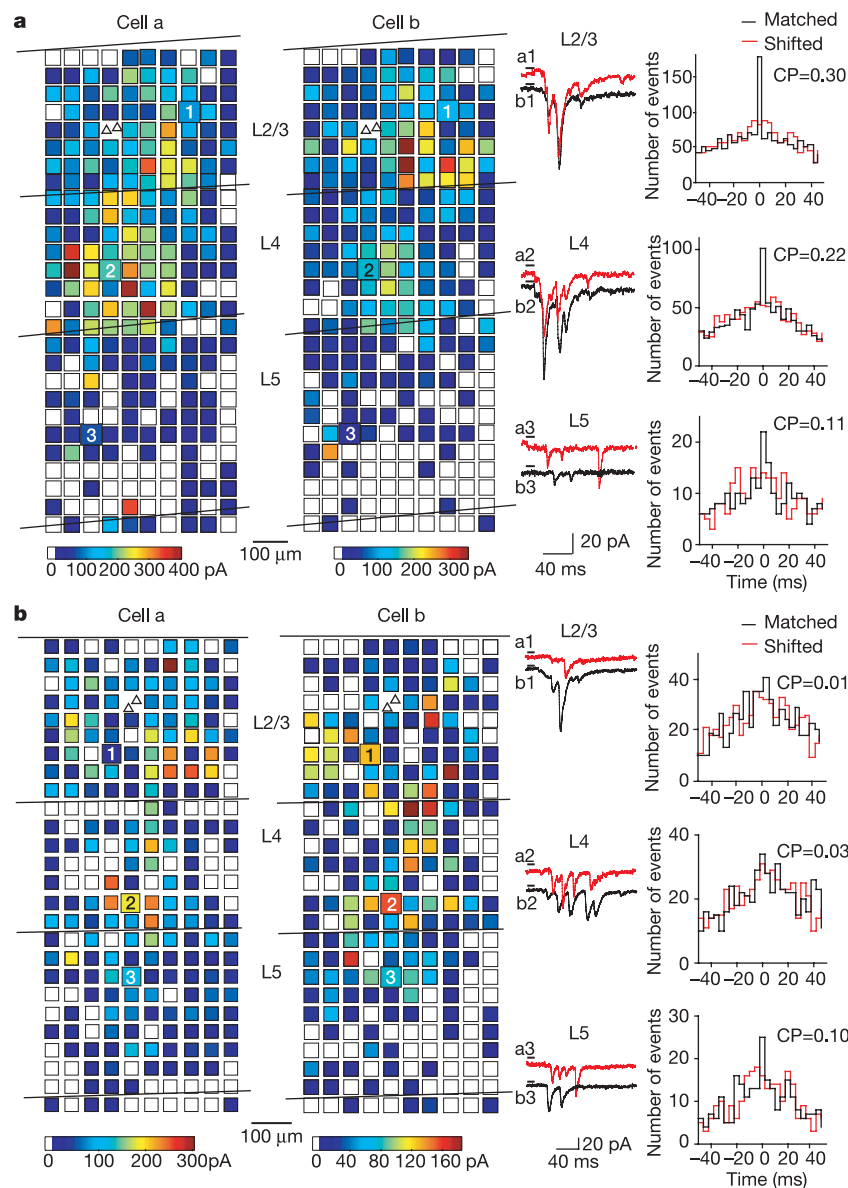


Figure 1 Cross-correlation analyses of photostimulation-evoked excitatory postsynaptic currents (EPSCs) simultaneously recorded in adjacent pairs of layer 2/3 pyramidal neurons. **a, b**, Results are shown for a pair of pyramidal neurons that was synaptically connected (**a**) and for a pair that was not connected (**b**). For each of the two cells, reconstructions of the locations of photostimulation sites (coloured squares) relative to the locations of laminar borders and cell bodies of recorded pyramidal neurons (triangles) are shown. The colour of each square indicates the sum of amplitudes of EPSCs that were observed in response to photostimulation at that site. To the right of these plots, example

voltage clamp recordings are shown for stimulation sites indicated by the large numbered squares. Simultaneous recordings from representative sites in each cortical layer are shown for 'cell a' (red) and 'cell b' (black). The short horizontal lines above each trace indicate the onset of photostimulation. The histograms to the far right of each panel show matched (black) and shifted (red) correlograms computed from data collected upon stimulation in each layer. The corresponding correlation probabilities (CPs) computed from these analyses are also indicated.

the network^{10–13}. Neighbouring neurons with extensively overlapping dendrites might share common input by chance, but information would be averaged across neurons to create a reliable but relatively uniform output. In the alternative scenario, there might be a fine-scale organization of connections between excitatory neurons within functional columns. The probability that two neurons are connected might be dependent on whether they share common input from other sources. This would reflect rules of connectivity that are not random but instead create substructure within each functional column. Fine-scale selectivity embedded within the columnar functional architecture might give rise to relatively independent neuronal networks that process information uniquely from their immediate neighbours.

To address these issues we took advantage of the ability of focal uncaging of glutamate ('photostimulation') to generate action potentials asynchronously in a small, spatially restricted population of neurons in rat visual cortex brain slices (see Supplementary Figs S1 and S2). By combining this type of stimulation with intracellular recordings of excitatory and inhibitory synaptic currents in pairs of adjacent layer 2/3 pyramidal neurons, we were able to use the timing of evoked synaptic currents to infer whether individual stimulated neurons provided common input to both recorded cells or if instead the recorded cells received input from separate neuronal populations. If a single presynaptic neuron is stimulated and it provides input to both recorded cells, it will generate synchronous synaptic currents; inputs from different presynaptic neurons that fire action potentials asynchronously will generate asynchronous synaptic currents.

We used established cross-correlation analysis methods¹⁴ to normalize for synchrony resulting from time-locking of action potential generation to the stimulus (see Methods). To obtain correlation probabilities, the numbers of synchronous synaptic currents attributable to shared input were expressed as a proportion of the total numbers of evoked synaptic currents from each cell (see Methods for details). The correlation probability closely estimates the probability that when a photostimulated presynaptic neuron fires an action potential and evokes a synaptic current in one of the two recorded layer 2/3 pyramidal neurons, the same presynaptic neuron will also evoke a synaptic current in the second recorded neuron. To determine the extent of shared input from the different laminar sources to each pair of recorded layer 2/3 pyramidal cells, separate calculations of correlation probability were made based on stimulation sites in each cortical layer. For example, for the pair of layer 2/3 pyramidal neurons illustrated in Fig. 1a, the correlation probability for stimulation sites in layer 4 was 0.22, meaning that for 22% of the cases in which a layer 4 neuron was stimulated and evoked an excitatory postsynaptic current (EPSC) in one layer 2/3 cell, that same layer 4 neuron also evoked a synchronous EPSC that was detected in the other recorded layer 2/3 cell. Similarly, the correlation probability of 0.30 obtained for this same pyramidal neuron pair for stimulation sites in layer 2/3 shows that these cells share about 30% of their excitatory layer 2/3 inputs.

To test whether connected pyramidal neurons belong to functional subnetworks, we compared correlation probabilities between connected and unconnected layer 2/3 pyramidal neuron pairs. We found that the correlation probabilities for excitatory connections from layer 4 and from layer 2/3 to layer 2/3 pyramidal neuron pairs depended on whether the recorded pyramidal neurons were connected to each other. When the recorded layer 2/3 pyramids were not connected, the correlation probabilities for EPSCs were very low (Figs 1b and 2a). For unconnected cell pairs, correlation probabilities averaged $3.8 \pm 1.1\%$ (mean $\pm$ s.e.m.) for layer 2/3 stimulation sites (range -5.3 to 10.3 , 17 cell pairs) and $3.6 \pm 0.9\%$ for layer 4 stimulation sites (range -2.2 to 9.1 , 17 cell pairs). In sharp contrast to the low correlation probability values for cell pairs that were not connected, correlation probabilities were high when layer 2/3 pyramids were connected (Figs 1a and 2a). For connected

cell pairs, correlation probabilities averaged $20.1 \pm 2.7\%$ for layer 2/3 stimulation sites (range 4.5 to 37.6, 16 cell pairs) and $16.8 \pm 2.1\%$ for layer 4 stimulation sites (range 2.9 to 30.0, 16 cell pairs). The differences in correlation probabilities for connected versus unconnected cell pairs were highly significant ($P < 0.0001$ for both layer 2/3 and layer 4 stimulation; Mann-Whitney U -test).

These data indicate that excitatory connections from layer 4 are highly selective at a fine scale, connecting preferentially to layer 2/3 pyramidal neurons that are in turn connected to each other. Furthermore, the fine-scale groupings of neurons that are defined by the selectivity of input from layer 4 are further reinforced by the fine-scale selectivity of excitatory connections within layer 2/3.

In contrast to excitatory connections from layer 4 and within layer 2/3, the specificity of excitatory connections from layer 5 was not dependent on whether the simultaneously recorded layer 2/3 pyramids were connected to each other. For stimulation sites in layer 5, correlation probabilities based on EPSCs averaged $9.8 \pm 1.7\%$ for unconnected cell pairs and $10.7 \pm 1.7\%$ for connected cell pairs (no significant difference, $P > 0.78$; Figs 1 and 2a). The finding that correlation probabilities do not differ between connected and unconnected cell pairs indicates that these connections link neurons across the fine-scale subnetworks established by the specificities of layer 4 and layer 2/3 connections.

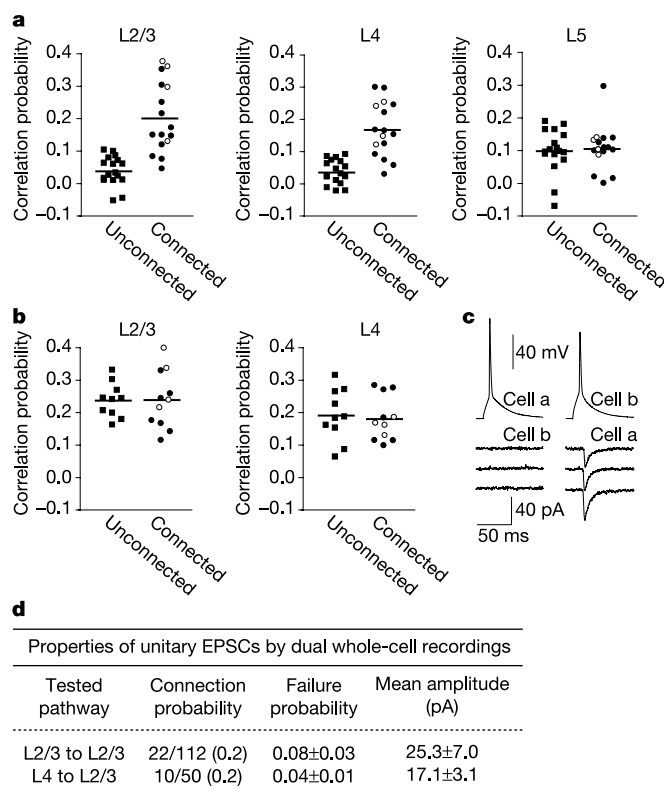


Figure 2 Correlation probabilities for EPSCs and IPSCs in layer 2/3 pyramidal cell pairs. **a, b**, Correlation probabilities for EPSCs (**a**) and IPSCs (**b**) are shown separately for photostimulation sites in each cortical layer and for connected (circles) versus unconnected (squares) pairs of cells. Open circles correspond to pyramidal cells that were reciprocally connected and filled circles indicate one-way connections. Mean values for each group are indicated by horizontal lines. **c, d**, An experiment testing connections between pyramidal cell pairs (**c**) and summary of results (**d**). Action potentials were triggered by injecting positive current into one current-clamped cell (upper traces) while recording EPSCs in the other voltage-clamped cell (lower traces, 3 examples shown for each). The cell pair illustrated in **c** had a one-way connection from 'cell b' to 'cell a'. Connections, when present, were highly reliable for both layer 4 to layer 2/3 (4% failure) and layer 2/3 to layer 2/3 (8% failure) pairs. Connections were found in 20% of cell pairs for both layer 4 to layer 2/3 and layer 2/3 to layer 2/3 pairs.

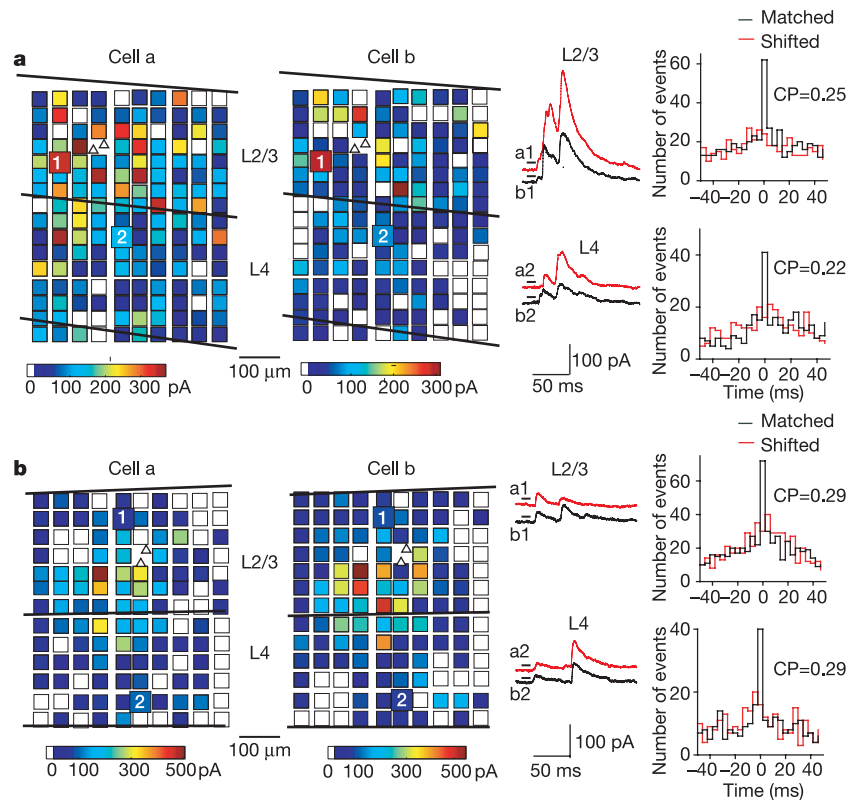


Figure 3 Cross-correlation analyses of photostimulation-evoked inhibitory postsynaptic currents (IPSCs) simultaneously recorded in adjacent pairs of layer 2/3 pyramidal neurons. See Fig. 1 legend for details.

The organization of inhibitory connections to adjacent layer 2/3 pyramidal cells was also independent from whether the layer 2/3 cells were connected. Therefore, inhibitory connections also do not respect the fine-scale groupings defined by the specificities of layer 4 and layer 2/3 excitatory connections. To estimate the extent of common inhibitory input onto adjacent pyramidal neurons from interneurons in each layer, the same photostimulation methods were employed except that the layer 2/3 pyramidal cells were voltage-clamped at 0 mV (see Methods) to allow recordings of inhibitory postsynaptic currents (IPSCs) as outward currents (Fig. 3). The correlation probabilities for IPSCs were typically high, regardless of whether the layer 2/3 pyramidal cells were connected and regardless of the layer that was stimulated (Figs 2b and 3). For connected cell pairs, correlation probabilities averaged $23.8 \pm 2.7\%$ (range 11.5 to 40.0) for stimulation sites in layer 2/3 and $18.0 \pm 6.8\%$ (range 9.8 to 28.4) for layer 4. For unconnected cell pairs the correlation probabilities averaged $23.8 \pm 1.7\%$ (range 16.2 to 33.1) for layer 2/3, and $19.1 \pm 2.5\%$ (range 6.3 to 31.6) for layer 4. There were no significant differences between connected cell pairs or between layers. Layer 5 is not included in this analysis because, as expected from previous studies¹⁵, photostimulation in layer 5 resulted in only modest or no significant increase in IPSCs above spontaneous IPSC levels (see Methods).

It is important to note that cortical inhibition comes from diverse cell types that make synapses to different parts of the dendritic arbours of pyramidal cells and for which connections might differ in their reliability and/or detectability¹⁶. It is therefore possible that the IPSCs we detect during photostimulation might preferentially represent inputs from certain inhibitory cell types. For example, IPSCs detected during photostimulation could be biased towards those from basket cells (which make strong synapses close to or at the cell body)¹⁶, relative to those from inhibitory cell types that make synapses at electrotonically distant sites¹⁶. Therefore, our results

might not generalize to all sources of inhibition of layer 2/3 pyramidal neurons.

In summary, the results presented here indicate that excitatory connections from layer 4 to layer 2/3 pyramidal cells and within layer 2/3 are highly specific on a fine scale, creating groups of selectively interconnected neurons (Fig. 4). Such groupings

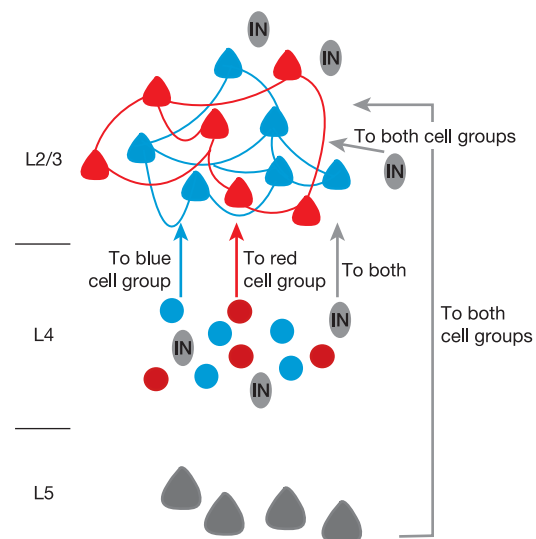


Figure 4 Schematic diagram illustrating the organization of cortical connections proposed in this study. Excitatory connections from layer 4 to layer 2/3 and within layer 2/3 define groups of selectively interconnected neurons (red or blue). The organization of excitation from layer 5 (grey triangles) and inhibition from layers 2/3 and 4 interneurons (IN, grey ovals) does not respect the fine-scale interconnected cell groups defined by excitatory connections from layer 4 and within layer 2/3.

are predicted from hebbian learning rules¹⁷, which are likely to regulate the development and maintenance of excitatory cortical connections^{18–20}.

Neither inhibitory connections nor excitatory connections from layer 5 provide shared input preferentially to connected layer 2/3 pyramidal cells (Fig. 4). Instead, these connections can serve to link neurons across the fine-scale subnetworks defined by excitatory connections from layers 2–4, and might act to modulate activity between the subnetworks. These observations do not, however, rule out the possibility that these connections might be specific for modes of organization not examined in these studies. For example there could be specificity related to an organization orthogonal to the fine-scale excitatory networks in layers 2–4, or on a different spatial scale.

Previous studies, using either photostimulation or more conventional methods, have demonstrated selectivity of connections to cortical neurons of distinctly different types^{15,21–28}. Here we have extended the photostimulation method to explicitly test for the possibility of fine-scale selectivity of cortical connections to neurons of the same type. We observe that the specificity of excitatory connections from layer 4 and within layer 2/3 depends on whether adjacent layer 2/3 pyramidal neurons are interconnected, independent of their locations (Fig. 4). These results therefore demonstrate selectivity of connections at a finer scale than cortical columnar architecture.

Like other species, the rat visual cortex has a columnar, retinotopic organization. But orientation-selective neurons in rat visual cortex are not organized into orientation columns; instead, adjacent neurons can be selective for disparate orientations²⁹. We therefore suggest that the fine-scale specificity of excitatory connections in rat visual cortex may be related to the emergence of orientation selectivity, similar to the columnar specificity found in cats and tree shrews^{3–6}. However, even in cat visual cortex where orientation columns are present, the excitatory connections from layer 4 to layer 3 are sparse; paired recordings show that layer 4 excitatory neurons connect to layer 3 pyramidal cells only 10% of the time⁹. And despite the presence of orientation columns, neighbouring neurons in cat visual cortex can differ in their selectivities for other features and sometimes demonstrate functional ‘micro-organization’³⁰. We therefore suggest that fine-scale selectivity of excitatory connections, embedded within the coarser columnar organization, is a common feature of cortical circuits. □

Methods

Slice preparation, photostimulation and recordings

The methods used for these experiments are similar to those reported previously¹⁵, with some differences as detailed here. A vibratome was used to cut 300 μm thick coronal brain slices from the primary visual cortex of P21–26 Long-Evans rats. Slices were cut in artificial cerebral spinal fluid (ACSF; 124 mM NaCl, 5 mM KCl, 1.25 mM KH_2PO_4 , 1.3 mM MgSO_4 , 3.2 mM CaCl_2 , 26 mM NaHCO_3 and 10 mM glucose) and stored in an interface chamber at $\sim 34^\circ\text{C}$ for at least one hour until they were transferred to a recording chamber containing ACSF with 60–80 μM ‘caged’ glutamate (γ -(α -carboxy-2-nitrobenzyl) ester, trifluoroacetate, L-glutamic acid) at room temperature; this is only half the concentration of caged glutamate used in previous studies¹⁵, and was used in order to reduce the numbers of neurons that fire action potentials synchronously (see Supplementary Fig. S2) and to reduce the size of direct responses so that postsynaptic currents (PSCs) can be detected upon stimulation at sites close to the recorded neurons. An infrared Olympus DIC microscope with a $\times 40$, 0.8 NA water immersion lens was used to visualize and target recording electrodes to pairs of layer 2/3 pyramidal neurons with somata separated by less than 50 μm for whole-cell recordings. The mean $\pm$ SEM distances between recorded cells were $34.6 \pm 3.9 \mu\text{m}$ for non-connected pairs and $35.1 \pm 4.1 \mu\text{m}$ for connected pairs of neurons. Cell bodies of recorded neurons were at least 50 μm from the surface of the slice. Glass recording electrodes (4–6 M Ω resistance) were filled with an intracellular solution consisting of 130 mM K-gluconate, 6 mM KCl, 2 mM MgCl_2 , 0.2 mM EGTA, 10 mM HEPES, 2.5 mM Na_2ATP , 0.5 mM Na_2GTP , 10 mM K-phosphocreatine and 0.3% biocytin, adjusted with KOH to pH 7.25. For some experiments in which IPSCs were recorded, potassium was replaced with cesium. All intracellular recordings had access resistances less than 20 M Ω . In all paired recordings, connections between neuron pairs were assessed by injecting current to evoke action potentials in one of the cells recorded in current-clamp while testing for PSCs during voltage-clamp recording in the other cell. For each pair, connections were tested in both directions for at least 50 trials, generating single action

potentials in each presynaptic neuron. When connections were not detected with this procedure, they were also tested by stimulation in trains of 4–5 action potentials at 50 Hz to induce possible potentiation of weak connections. Control experiments measuring spatial and temporal properties of photostimulation-evoked action potentials (see Supplementary Information) used extracellular loose-patch recordings made with the same recording electrodes, except that they were filled with ACSF.

Photostimulation was achieved by uncaging glutamate with 10 ms flashes of ultraviolet light from an argon-ion laser focused through the $\times 40$ microscope objective. This results in the generation of action potentials only in neurons with cell bodies within 100 μm (and usually less than 50 μm) of the site of uncaging (see Supplementary Fig. S1). Photostimulation-evoked synaptic currents were measured from voltage-clamped neurons, with the holding potentials at -65 mV to measure EPSCs and at 0 mV to measure IPSCs. Spontaneous synaptic currents were also recorded in interleaved trials with no stimulation.

Data analysis

Maps of photostimulation sites were aligned to laminar borders in fixed and stained tissue¹⁵ (for example, Figs 1 and 3) and each site was assigned a laminar identity. Sites within 50 μm of laminar borders were discarded from further analyses in order to limit the number of evoked synaptic currents arising from neurons with cell bodies potentially outside the stimulated layer. The electrical recordings from photostimulation and no-stimulation (control) trials were analysed using peak analysis software from Synaptosoft and other custom software. The times of onset and amplitudes of all EPSCs or IPSCs occurring within 150 ms of stimulation were marked. Rise times of PSCs were measured as the time taken for the amplitude to increase from 10% to 90% of its peak value. Results from analyses of the laminar sources and strengths of excitatory and inhibitory input to layer 2/3 pyramidal neurons (not shown) were indistinguishable from those described previously¹⁵, and there were no systematic differences that correlated with results from cross-correlation analyses.

Cross-correlograms of EPSCs and/or IPSCs were computed for each pair of simultaneously recorded layer 2/3 pyramidal neurons; separate correlograms were computed for stimulation sites from each cortical layer (layers 2/3, 4 and 5 for EPSCs and layers 2/3 and 4 for IPSCs). Other layers provided weak or variable input to recorded neurons, preventing evoked PSCs from being clearly distinguished from spontaneous PSCs. Correlograms were also computed for spontaneous PSCs. Cross-correlation data were binned into histograms using 4 ms bins; the central bin included values of $0 \pm 2 \text{ ms}$. Data from the stimulation trials (from the same layer) were also used to create shifted correlograms for each layer and cell pair¹⁴. To calculate the correlation probability, the shifted correlogram was subtracted from the unshifted correlogram for the corresponding layer, and then the value in the central bin was divided by the average estimated total number of evoked PSCs (for the two cells) observed for all trials in the relevant layer. The average number of evoked PSCs was calculated as the total number of measured PSCs for ‘cell a’ minus the expected number of spontaneous PSCs for that cell, plus the same value calculated for ‘cell b’, divided by 2.

Received 9 September; accepted 6 December 2004; doi:10.1038/nature03252.

- Mountcastle, V. B. Introduction. *Computation in cortical columns. Cereb. Cortex* **13**, 2–4 (2003).
- Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154 (1962).
- Chapman, B., Zahs, K. R. & Stryker, M. P. Relation of cortical cell orientation selectivity to alignment of receptive fields of the geniculocortical afferents that arborize within a single orientation column in ferret visual cortex. *J. Neurosci.* **11**, 1347–1358 (1991).
- Ferster, D., Chung, S. & Wheat, H. Orientation selectivity of thalamic input to simple cells of cat visual cortex. *Nature* **380**, 249–252 (1996).
- Alonso, J. M., Usrey, W. M. & Reid, R. C. Rules of connectivity between geniculate cells and simple cells in cat primary visual cortex. *J. Neurosci.* **21**, 4002–4015 (2001).
- Mooser, F., Bosking, W. H. & Fitzpatrick, D. A morphological basis for orientation tuning in primary visual cortex. *Nature Neurosci.* **7**, 872–879 (2004).
- Callaway, E. M. Local circuits in primary visual cortex of the macaque monkey. *Annu. Rev. Neurosci.* **21**, 47–74 (1998).
- Thomson, A. M. & Morris, O. T. Selectivity in the inter-laminar connections made by neocortical neurones. *J. Neurocytol.* **31**, 239–246 (2002).
- Thomson, A. M., West, D. C., Wang, Y. & Bannister, A. P. Synaptic connections and small circuits involving excitatory and inhibitory neurons in layers 2–5 of adult rat and cat neocortex: triple intracellular recordings and biocytin labelling *in vitro*. *Cereb. Cortex* **12**, 936–953 (2002).
- Hellwig, B. A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biol. Cybern.* **82**, 111–121 (2000).
- Hellwig, B., Schuz, A. & Aertsen, A. Synapses on axon collaterals of pyramidal cells are spaced at random intervals: a Golgi study in the mouse cerebral cortex. *Biol. Cybern.* **71**, 1–12 (1994).
- Braitenberg, V. & Schuz, A. *Anatomy of the Cortex* (Springer-Verlag, Berlin, 1991).
- Binzegger, T., Douglas, R. J. & Martin, K. A. A quantitative map of the circuit of cat primary visual cortex. *J. Neurosci.* **24**, 8441–8453 (2004).
- Aertsen, A. M., Gerstein, G. L., Habib, M. K. & Palm, G. Dynamics of neuronal firing correlation: modulation of ‘effective connectivity’. *J. Neurophysiol.* **61**, 900–917 (1989).
- Dantzker, J. L. & Callaway, E. M. Laminar sources of synaptic input to cortical inhibitory interneurons and pyramidal neurons. *Nature Neurosci.* **3**, 701–707 (2000).
- Kawaguchi, Y. & Kondo, S. Parvalbumin, somatostatin and cholecystokinin as chemical markers for specific GABAergic interneuron types in the rat frontal cortex. *J. Neurocytol.* **31**, 277–287 (2002).
- Hebb, D. O. *The Organization of Behavior* (Wiley, New York, 1949).
- Yoshimura, Y., Ohmura, T. & Komatsu, Y. Two forms of synaptic plasticity with distinct dependence on age, experience, and NMDA receptor subtype in rat visual cortex. *J. Neurosci.* **23**, 6557–6566 (2003).
- Sur, M., Schummers, J. & Dragoi, V. Cortical plasticity: time for a change. *Curr. Biol.* **12**, R168–R170 (2002).

20. Katz, L. C. & Shatz, C. J. Synaptic activity and the construction of cortical circuits. *Science* **274**, 1133–1138 (1996).
21. Sawatari, A. & Callaway, E. M. Diversity and cell type specificity of local excitatory connections to neurons in layer 3B of monkey primary visual cortex. *Neuron* **25**, 459–471 (2000).
22. Schubert, D., Kotter, R., Zilles, K., Luhmann, H. J. & Staiger, J. F. Cell type-specific circuits of cortical layer IV spiny neurons. *J. Neurosci.* **23**, 2961–2970 (2003).
23. Agmon, A. & Connors, B. W. Correlation between intrinsic firing patterns and thalamocortical synaptic responses of neurons in mouse barrel cortex. *J. Neurosci.* **12**, 319–329 (1992).
24. Gibson, J. R., Beierlein, M. & Connors, B. W. Two networks of electrically coupled inhibitory neurons in neocortex. *Nature* **402**, 75–79 (1999).
25. Gonchar, Y. & Burkhalter, A. Connectivity of GABAergic calretinin-immunoreactive neurons in rat primary visual cortex. *Cereb. Cortex* **9**, 683–696 (1999).
26. Gonchar, Y. & Burkhalter, A. Distinct GABAergic targets of feedforward and feedback connections between lower and higher areas of rat visual cortex. *J. Neurosci.* **23**, 10904–10912 (2003).
27. Meskenaite, V. Calretinin-immunoreactive local circuit neurons in area 17 of the cynomolgus monkey, *Macaca fascicularis*. *J. Comp. Neurol.* **379**, 113–132 (1997).
28. Staiger, J. F. *et al.* Innervation of interneurons immunoreactive for VIP by intrinsically bursting pyramidal cells and fast-spiking interneurons in infragranular layers of juvenile rat neocortex. *Eur. J. Neurosci.* **16**, 11–20 (2002).
29. Girman, S. V., Sauve, Y. & Lund, R. D. Receptive field properties of single neurons in rat primary visual cortex. *J. Neurophysiol.* **82**, 301–311 (1999).
30. DeAngelis, G. C., Ghose, G. M., Ohzawa, I. & Freeman, R. D. Functional micro-organization of primary visual cortex: receptive field analysis of nearby neurons. *J. Neurosci.* **19**, 4046–4064 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful for support from the National Institutes of Health. We thank Y. Komatsu and F. Briggs and members of the Callaway laboratory for discussions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.M.C. (callaway@salk.edu).

Different time courses of learning-related activity in the prefrontal cortex and striatum

Anitha Pasupathy & Earl K. Miller

The Picower Center for Learning and Memory, RIKEN-MIT Neuroscience Research Center and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA

To navigate our complex world, our brains have evolved a sophisticated ability to quickly learn arbitrary rules such as ‘stop at red’. Studies in monkeys using a laboratory test of this capacity—conditional association learning—have revealed that frontal lobe structures (including the prefrontal cortex) as well as subcortical nuclei of the basal ganglia are involved in such learning^{1–5}. Neural correlates of associative learning have been observed in both brain regions^{6–14}, but whether or not these regions have unique functions is unclear, as they have typically been studied separately using different tasks. Here we show that during associative learning in monkeys, neural activity in these areas changes at different rates: the striatum (an input structure of the basal ganglia) showed rapid, almost bistable, changes compared with a slower trend in the prefrontal cortex that was more in accordance with slow improvements in behavioural performance. Also, pre-saccadic activity began progressively earlier in the striatum but not in the prefrontal cortex as learning took place. These results support the hypothesis that rewarded associations are first identified by the basal ganglia, the output of which ‘trains’ slower learning mechanisms in the frontal cortex¹⁵.

The prefrontal cortex (PFC) is a cortical area important for the organization of goal-directed, rule-based behaviours; the basal ganglia are a group of subcortical nuclei long associated with the

control of volitional movements^{1–3,16–18}. Both of these areas receive inputs from many brain systems (for example, sensory, motor and reward), which makes them well suited for roles in learning. Their anatomy also suggests a close relationship—the PFC and basal ganglia are interconnected in cortico-basal ganglionic ‘loops’^{19,20}—but the nature of this interaction is still unclear. Some results have led to the suggestion of a sequential relationship, in which the PFC is involved in new learning and the basal ganglia are subsequently involved in consolidating familiar routines into automatic habits^{21,22}. Another hypothesis, not necessarily incompatible with the one above, suggests a dominant role for the basal ganglia in new learning^{15,23} due to its anatomical architecture and the membrane properties of striatal spiny neurons. These hypotheses lead to specific predictions about the time course of learning in these areas: based on the first hypothesis, the PFC is predicted to lead the basal ganglia; based on the second hypothesis, the basal ganglia lead the PFC. Here, we report evidence in favour of the latter event; that is, learning-related changes appear sooner and progress more rapidly in the striatum than the PFC.

To test these hypotheses, we simultaneously recorded neural activity from the dorsolateral PFC (areas 9 and 46) and the head and body of the caudate nucleus, a part of the striatum that receives direct projections from, and indirectly projects to, the PFC^{19,20} (see Methods). Monkeys learned associations between each of two visual cues and two saccadic eye movements (right and left, Fig. 1a). Monkeys were familiar with the task, but each day two novel cues were used and their associations learned by trial and error using juice reward as feedback. Once the cue–saccade associations had been learned, they were reversed without warning and the opposite pairing was then learned (see Supplementary Note 1).

Figure 1b (left) shows the average behavioural performance before and after the reversals. Saccade choices dropped to about 0% correct for the first few trials after the reversal because the previous associations were still being followed. Then, performance jumped to chance (50%) followed by a slow increase with trial number. Likewise, reaction time increased by an average of about

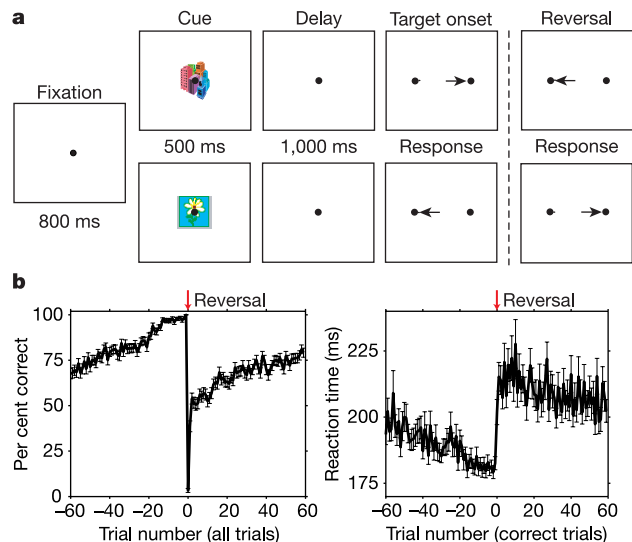


Figure 1 Task and behaviour. **a**, One of two initially novel cues was briefly presented at centre of gaze followed by a memory delay and then presentation of two target spots on the right and left. Saccade to the target associated with the cue at that time was rewarded. After this was learned, the cue–saccade associations were reversed and re-learned. **b**, Average per cent correct performance (left) and reaction time (right) across sessions and blocks as a function of trial number (left: all trials; right: correct trials only) during learning for two monkeys. Zero (indicated by red arrow) represents first trial after reversal. Error bars show standard error of the mean.

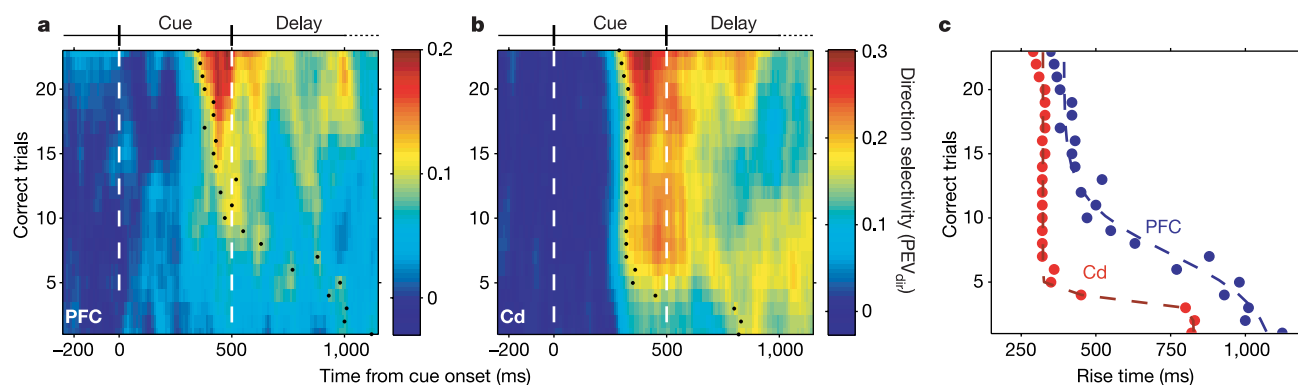


Figure 2 Change in peri-cue saccade direction selectivity in prefrontal cortex and caudate nucleus with learning. **a,b**, Population strength of direction selectivity (PEV_{dir}: proportion of explainable variance by direction factor) (colour scale) shown as a function of correct trials and time from cue onset for PFC (**a**) and caudate nucleus (Cd) (**b**) during cue (white lines) and delay periods. Black dots indicate 'rise time' (time to half-maximum selectivity).

Selectivity strength increases and appears earlier in both areas as learning takes place. Changes appear earlier and reach an asymptote sooner in the caudate nucleus than the PFC. **c**, Rise times for PFC (blue) and Cd (red). Dotted lines show sigmoids of best fit. Data shown in **a–c** are based on correct trials collapsed across all blocks (reversals).

50 ms in the first few correct trials after reversal and then gradually decreased as the reversed associations were learned (Fig. 1b (right), see Supplementary Note 2). Because the monkeys did not instantly reverse the associations, we could examine learning across multiple trial blocks in each recording session. Across 51 sessions, we examined the activity of 432 PFC and 279 caudate nucleus neurons (see Methods).

Many PFC and caudate nucleus neurons showed activity that reflected the saccade direction (PFC: 39% or 168/432; caudate nucleus: 36% or 101/279 of all recorded neurons), especially around the time of its execution. Neuronal activity also reflected the cues or their associations with the saccades (see Supplementary Note 3). As the monkeys learned which associations would yield reward, there was an increase in early-trial activity (that is, activity around the time of cue presentation) that predicted the direction of the saccade to be made after the delay. Examples of single PFC and caudate nucleus neurons with this 'prospective' activity and the development of this activity with learning are shown in Supplementary Fig. 1.

We assessed learning-related changes in saccade direction selectivity for the 168 PFC and 101 caudate nucleus neurons showing such selectivity during any trial period (analysis of variance (ANOVA), $P < 0.01$). Selectivity was quantified with a regression

analysis that measured the proportion of explainable variance in activity accounted for by saccade direction (PEV_{dir}; see Methods). PEV_{dir} is shown for the PFC and caudate nucleus populations as a function of time during cue and delay periods and number of correct trials (Fig. 2a, b). During the first few correct trials early in the learning process, both populations showed relatively weak early-trial direction selectivity. This strength of selectivity increased with the number of correct trials (especially near the end of cue presentation), albeit at different rates in the PFC and caudate nucleus neurons: selectivity increased sooner and more abruptly in the caudate nucleus compared with the PFC (see Supplementary Note 4).

This faster increase in early-trial direction selectivity in the caudate nucleus can be seen in Fig. 2c, which shows the time when half-maximum selectivity was reached (the 'rise time') for each neuron population on each trial. In the first few correct trials, rise time is late in the delay, near the time of saccade execution. After just a few correct trials, rise time in the caudate nucleus is much earlier as the strength of 'prospective' direction selectivity rapidly increases in the cue period (see Supplementary Note 5). In contrast, the more gradual increase of early-trial direction selectivity in the PFC results in a slower leftward shift in rise time. Best-fitting sigmoidal curves (dashed lines) confirmed that the shift in rise

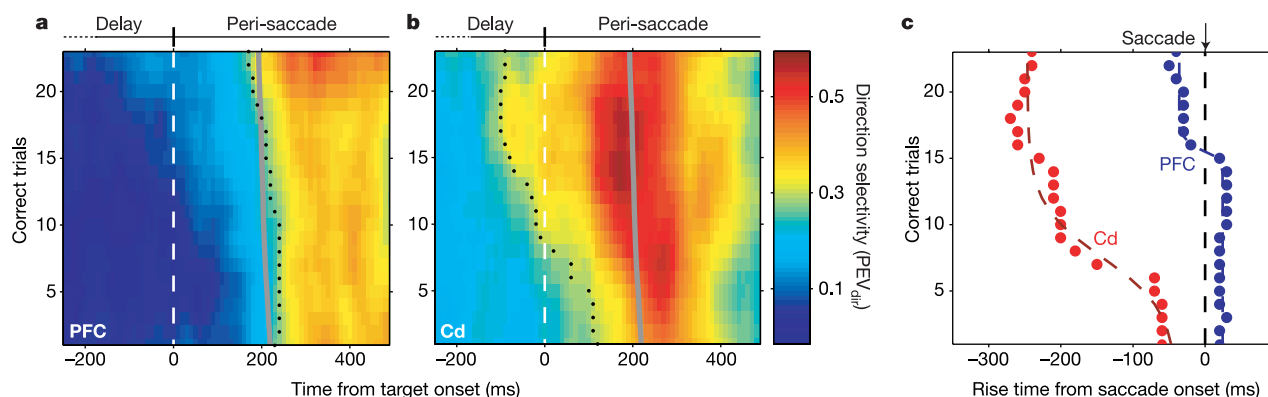


Figure 3 Change in saccade direction selectivity at the time of saccade execution during the learning process. The same neuron populations and conventions were used as in Fig. 2. **a, b**, White dashed lines denote the onset of target spots. Grey lines represent average reaction time. Peak direction selectivity (colour scale) in PFC (**a**) and Cd (**b**) remains relatively constant, but appears progressively earlier with learning in the caudate

nucleus, as illustrated by rise time (black dots). Note that the colour scales here are higher than in Fig. 2 (peri-cue period) because direction selectivity is much stronger during saccade execution. **c**, Rise times for the PFC (blue) and Cd (red) populations relative to saccade onset. Black dashed line represents time of saccade onset.

time was faster and reached an asymptote sooner in the caudate nucleus (maximum slope (at trial 4) = 480 ms per trial, asymptote at trial 6) than the PFC (maximum slope (at trial 7) = 108 ms per trial, asymptote at trial 12) (see Supplementary Note 6). This sudden increase in early-trial direction selectivity after only a few correct trials could even be seen in some single neurons from the caudate nucleus (see Supplementary Fig. 2). This is in sharp contrast to the monkeys' more gradual improvement in performance: 95% of the change in correct choices and reaction time was only achieved after 20 correct trials per cue. Indeed, PFC rise times showed a significantly stronger correlation with per cent correct performance (linear correlation coefficient $r = -0.96$) than caudate nucleus rise times ($r = -0.79$) (see Supplementary Notes 7 and 8).

Figure 3a, b shows the average saccade direction selectivity for the same neuron populations before and after saccade execution. Here, both areas showed strong selectivity, but with clear differences. For each trial, rise time and peak were earlier in the caudate nucleus than in the PFC. Additionally, with increasing numbers of correct trials, selectivity became increasingly pre-saccadic in the caudate nucleus but not in the PFC; it began and peaked progressively earlier relative to target onset ('go' signal) (Fig. 3a, b) and saccade initiation (Fig. 3c). In the first few correct trials, rise time in the caudate nucleus was about 60 ms before the saccade, but occurred 250 ms before the saccade after 20 correct trials (Fig. 3c). In contrast, PFC rise times were relatively stable during learning and were centred closely around saccade initiation.

These results illustrate differences between PFC and caudate nucleus activity during conditional association learning. Early-trial caudate nucleus activity quickly reflected the forthcoming saccade, whereas such activity appeared more slowly in the PFC. This fits with observations of a striatal infrastructure ideal for rapid, supervised (reward-based) learning^{24,25}. One possibility is that the PFC and caudate nucleus are components of different learning systems that are set in opposition in our task. The cue-response associations may have engaged the caudate nucleus; the striatum is thought to be central to the 'habit memory' that establishes such links. However, flexibility (for example, reversal) has been associated with the PFC, so perhaps activity in the caudate nucleus was 'ignored' in favour of PFC mechanisms with slower plasticity, but consequently greater flexibility. Also, our results may support hypotheses that learning in the frontal cortex could be 'trained' by the basal ganglia^{15,26}. Dopaminergic reward-prediction error signals from the midbrain^{27,28} may allow rapid formation of reward-relevant associations in the striatum²³, which over a course of trials might train slower, and more graded, hebbian mechanisms in the PFC via the output nuclei of the basal ganglia and the thalamus¹⁵ (see Supplementary Note 9). Behaviour may follow changes in the PFC or a combination of the PFC and striatum (and other areas²⁹); this would explain the overall slower time course of behavioural improvement relative to changes in the caudate nucleus. Although pre-saccadic activity might also be linked to the improvement in choices with learning, it seems likely to reflect the decrease in reaction time; a correlation between caudate nucleus (and frontal cortex) pre-saccadic activity and reaction time has been previously demonstrated^{10,30}. These results indicate that during conditional visuomotor learning, changes in caudate nucleus activity can lead those in the PFC. □

Methods

Behavioural task

Each trial began with the presentation of a fixation spot, followed by 500 ms of cue presentation, and 1000 ms of memory delay. Monkeys were required to maintain gaze within 1.5° of the fixation spot during these periods. After the delay, the fixation spot was extinguished and two targets appeared on the right and left. A direct saccade to the target associated with the cue yielded reward (see Supplementary Note 10). After performance reached criterion ($\geq 90\%$ correct over 10 trials per cue) and there were at least 30 correct trials for each cue, the associations were reversed. Monkeys completed three to eight

reversals (four to nine trial blocks) per recording session (average six trial blocks). For each session, two new cues (complex, multi-coloured images) were selected at random. In addition, there were two highly familiar, unchanging cue-response associations that were randomly intermingled and presented half as often. Results using these familiar cues will be reported in future publications.

Data collection

Neural activity was recorded from the dorsolateral prefrontal cortex (areas 9 and 46) and the head and body of the caudate nucleus. Recording wells were positioned stereotactically based on images obtained using magnetic resonance imaging. All animal procedures conformed to NIH guidelines and the MIT Committee on Animal Care. Arrays of 12–24 (8–16 in the PFC and 4–8 in the caudate nucleus) dura-puncturing tungsten microelectrodes (FHC Instruments), were mounted on custom-made, independently adjustable microdrives. All isolated neurons were accepted for study without being pre-screened. Waveforms were digitized, stored and sorted offline based on waveform shape characteristics.

Data analysis

Two-way balanced analyses of variance (ANOVAs) were conducted on the average neuronal activity during each of four periods: 'cue' (500 ms, starting 100 ms after cue onset) 'delay' (end of cue period to 150 ms before saccade onset), 'saccade' (300 ms centred on saccade onset) and 'reward' (250 ms, starting 50 ms after reward onset). As in previous work⁷, selectivity after learning was based on the last 10 correct trials per association before reversal in each block, but results were similar when selectivity was based on all correct trials. All tests were evaluated at $P < 0.01$.

Saccade direction selectivity was quantified as the fraction of each neuron's variance explained by saccade direction. The total variance (σ^2) was partitioned into object (σ_{obj}^2), direction (σ_{dir}^2), interaction (σ_{int}^2) and error (σ_{err}^2) terms. Selectivity strength (R^2 for the direction factor) was quantified as ($\sigma_{\text{dir}}^2/\sigma^2$). The proportion of total explainable variance was ($1 - \sigma_{\text{err}}^2/\sigma^2$). R^2 was computed for each neuron over a 100 ms centred window, slid in 10 ms steps throughout the trial (see Supplementary Note 11). To quantify changes across trials, R^2 was calculated for each neuron across an eight-trial window, slid in one-trial steps over the first 30 correct trials per cue per trial block, collapsed across blocks. Analysis was restricted to the first 30 correct trials per association, the minimum block length, because block length varied with learning rate.

We compared direction selectivity across the PFC and caudate nucleus populations (Figs 2 and 3) by computing the proportion of explainable variance accounted for by the direction factor (PEV_{dir}) as the ratio of average R^2 and the average total explainable variance across cells. Thus PEV_{dir} represents the population strength of direction selectivity. Using R^2 instead of PEV_{dir} yielded similar results, but PEV_{dir} is advantageous because it expresses saccade direction selectivity as a proportion of explainable variance without including unexplained variance (that is, that due to uncontrolled variables and noise). To assess the trend in direction selectivity with learning, we determined the half-maximum (across all trials) PEV_{dir} for each population separately and then calculated rise time as the time at which PEV_{dir} in each trial reached that value (see Supplementary Note 12).

Received 31 August; accepted 17 December 2004; doi:10.1038/nature03287.

- Petrides, M. In *Handbook of Neuropsychology* (eds Boller, F. & Grafman, J.) 59–82 (Elsevier, Amsterdam, 1994).
- Passingham, R. E. *The Frontal Lobes and Voluntary Action* (Oxford Univ. Press, Oxford, 1995).
- Fuster, J. M. *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe* (Lippincott-Raven, Philadelphia, 1997).
- Wise, S. P., Murray, E. A. & Gerfen, C. R. The frontal cortex-basal ganglia system in primates. *Crit. Rev. Neurobiol.* **10**, 317–356 (1996).
- Murray, E. A., Bussey, T. J. & Wise, S. P. Role of prefrontal cortex in a network for arbitrary visuomotor mapping. *Exp. Brain Res.* **133**, 114–129 (2000).
- Tremblay, L., Hollerman, J. R. & Schultz, W. Modifications of reward expectation-related neuronal activity during learning in primate striatum. *J. Neurophysiol.* **80**, 964–977 (1998).
- Asaad, W. F., Rainer, G. & Miller, E. K. Neural activity in the primate prefrontal cortex during associative learning. *Neuron* **21**, 1399–1407 (1998).
- White, I. M. & Wise, S. P. Rule-dependent neuronal activity in the prefrontal cortex. *Exp. Brain Res.* **126**, 315–335 (1999).
- Toni, I. & Passingham, R. E. Prefrontal-basal ganglia pathways are involved in the learning of arbitrary visuomotor associations: a PET study. *Exp. Brain Res.* **127**, 19–32 (1999).
- Lauwereyns, J., Watanabe, K., Coe, B. & Hikosaka, O. A neural correlate of response bias in monkey caudate nucleus. *Nature* **418**, 413–417 (2002).
- Hadj-Bouziane, F. & Boussaoud, D. Neuronal activity in the monkey striatum during conditional visuomotor learning. *Exp. Brain Res.* **153**, 190–196 (2003).
- Schumacher, E. H., Elston, P. A. & D'Esposito, M. Neural evidence for representation-specific response selection. *J. Cogn. Neurosci.* **15**, 1111–1121 (2003).
- Brasted, P. J. & Wise, S. P. Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum. *Eur. J. Neurosci.* **19**, 721–740 (2004).
- Hoshi, E. & Tanji, J. Area-selective neuronal activity in the dorsolateral prefrontal cortex for information retrieval and action planning. *J. Neurophysiol.* **91**, 2707–2722 (2004).
- Houk, J. C. & Wise, S. P. Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. *Cereb. Cortex* **5**, 95–110 (1995).
- Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
- Graybiel, A. M. The basal ganglia and the initiation of movement. *Rev. Neurol. (Paris)* **146**, 570–574 (1990).
- Hikosaka, O., Takikawa, Y. & Kawagoe, R. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol. Rev.* **80**, 953–978 (2000).

19. DeLong, M. R. & Georgopoulos, A. P. in *Handbook of Physiology—Nervous System* (eds Brookhart, J. M. & Mountcastle, V. B.) 1017–1061 (American Physiological Society, Bethesda, 1981).
20. Middleton, F. A. & Strick, P. L. Basal-ganglia 'projections' to the prefrontal cortex of the primate. *Cereb. Cortex* **12**, 926–935 (2002).
21. Packard, M. G. & Knowlton, B. J. Learning and memory functions of the basal ganglia. *Annu. Rev. Neurosci.* **25**, 563–593 (2002).
22. Graybiel, A. M. The basal ganglia and chunking of action repertoires. *Neurobiol. Learn. Mem.* **70**, 119–136 (1998).
23. Bar-Gad, I., Morris, G. & Bergman, H. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Prog. Neurobiol.* **71**, 439–473 (2003).
24. Reynolds, J. N., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413**, 67–70 (2001).
25. Wilson, C. J. & Kawaguchi, Y. The origins of two-state spontaneous membrane potential fluctuations of neostriatal spiny neurons. *J. Neurosci.* **16**, 2397–2410 (1996).
26. O'Reilly, R. C. & Munakata, Y. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Stimulating the Brain* (MIT Press, Cambridge, Massachusetts, 2000).
27. Schultz, W. & Dickinson, A. Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* **23**, 473–500 (2000).
28. McClure, S. M., Berns, G. S. & Montague, P. R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339–346 (2003).
29. Wirth, S. *et al.* Single neurons in the monkey hippocampus and learning of new associations. *Science* **300**, 1578–1581 (2003).
30. Hanes, D. P. & Schall, J. D. Neural control of voluntary movement initiation. *Science* **274**, 427–430 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. H. Histed for valuable discussions; K. J. MacCully for technical assistance; W. F. Asaad, A. J. Bastian, T. Buschman, A. C. Diogo, J. Feingold, D. J. Freedman, M. Machon, J. McDermott, J. E. Roy and M. Wicherski for helpful comments. This work was supported by a grant from the N.I.N.D.S. A.P. was supported by the Tourette's Syndrome Association.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.P. (anitha@mit.edu).

CFTR channel opening by ATP-driven tight dimerization of its nucleotide-binding domains

Paola Vergani¹, Steve W. Lockless², Angus C. Nairn^{3,4} & David C. Gadsby¹

¹Laboratory of Cardiac/Membrane Physiology, ²Laboratory of Molecular Neurobiology and Biophysics, and ³Laboratory of Molecular and Cellular Neuroscience, The Rockefeller University, New York, New York 10021, USA
⁴Department of Psychiatry, Yale University, New Haven, Connecticut 06519, USA

ABC (ATP-binding cassette) proteins constitute a large family of membrane proteins that actively transport a broad range of substrates. Cystic fibrosis transmembrane conductance regulator (CFTR), the protein dysfunctional in cystic fibrosis, is unique among ABC proteins in that its transmembrane domains comprise an ion channel. Opening and closing of the pore have been linked to ATP binding and hydrolysis at CFTR's two nucleotide-binding domains, NBD1 and NBD2 (see, for example, refs 1, 2). Isolated NBDs of prokaryotic ABC proteins dimerize upon binding ATP, and hydrolysis of the ATP causes dimer dissociation^{3–5}. Here, using single-channel recording methods on intact CFTR molecules, we directly follow opening and closing of the channel gates, and relate these occurrences to ATP-mediated events in the NBDs. We find that energetic coupling⁶ between two CFTR residues, expected to lie on opposite sides of its predicted NBD1–NBD2 dimer interface, changes in concert with channel gating status. The two monitored side chains are independent of each other in closed channels but become coupled as the channels open. The results directly link ATP-driven tight dimerization of

CFTR's cytoplasmic nucleotide-binding domains to opening of the ion channel in the transmembrane domains. This establishes a molecular mechanism, involving dynamic restructuring of the NBD dimer interface, that is probably common to all members of the ABC protein superfamily.

Crystal structures of most ABC-protein NBDs determined so far share the same fold^{7,8} with a core subdomain ('head') that binds the ATP, and an α -helical subdomain ('tail') that includes the ABC-specific signature sequence (LSGGQ). Dimeric structures revealed nucleotide-bound NBD homodimers in rotationally symmetric 'head-to-tail' arrangement, enclosing two ATP molecules within interfacial composite sites, each comprising conserved ATP-binding motifs from the head of one monomer and signature sequence residues from the tail of the other^{3,5,9,10}. On the basis of this structural evidence and biochemical studies of reversible dimerization of isolated NBDs^{4,5,11–13}, opening and closing of CFTR channels can be interpreted¹⁴ in terms of cycles of NBD1–NBD2 dimerization and dissociation, induced by ATP binding and hydrolysis, respectively (Fig. 1a). Opening of a phosphorylated CFTR Cl^- channel seems to require ATP binding to both composite sites because, at low [ATP], mutations expected to weaken ATP binding can make nucleotide occupancy at either site rate-limiting for channel opening¹⁴. In addition, interfering with hydrolysis prevents the normal rapid closing of CFTR channels^{1,2,14}. Because photolabelling studies show that ATP can remain at the NBD1-head site for several minutes without being hydrolysed^{15,16}, whereas a CFTR-channel gating cycle lasts only seconds, channel opening and closing seem to be timed by nucleotide binding and hydrolysis at the composite site

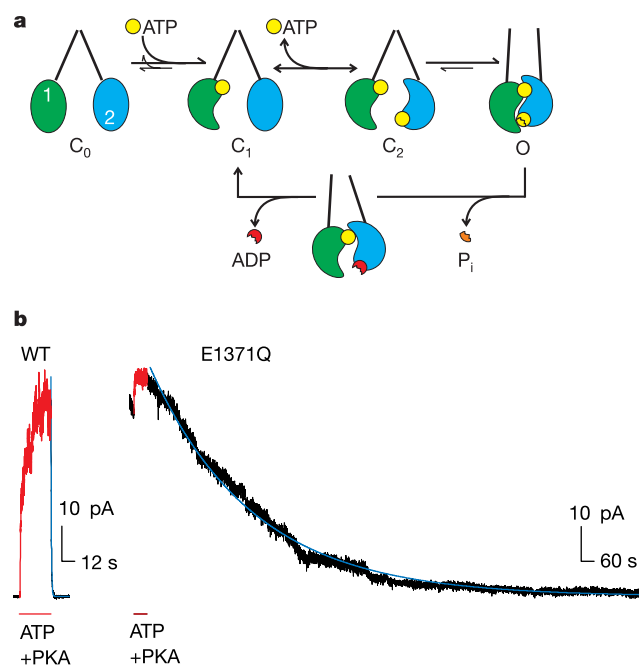


Figure 1 Open CFTR channels correspond to dimerized NBDs. **a**, Diagram illustrating the proposed mechanism coupling the opening of the Cl^- channel pore (C_m , closed states; O, open) in the transmembrane domains (converging, or semi-parallel, straight lines) to the hydrolysis cycle through the dimerization of NBDs (green, NBD1; blue, NBD2). The dynamic formation and disruption of a tight NBD dimer interface are represented by major changes in shape and position simply for clarity (see text). **b**, Mutating the 'Walker B' glutamate, Glu 1371, in NBD2 markedly increases the stability of the Cl^- channel's open burst state. Records from patches containing hundreds of channels, activated by exposure to 5 mM ATP and 300 nM cAMP-dependent protein kinase (PKA, red). Time constants for current decay fit lines (blue): WT, $\tau = 0.45$ s; E1371Q, $\tau = 476$ s. Note the fivefold expanded timescale for the WT record.

incorporating the NBD2 head, which we call the NBD2 catalytic site^{14,15}.

Evidence supporting our speculation (Fig. 1a) that the CFTR-channel open state corresponds to the dimerized NBD conformation is provided by the approximately 1,000-fold stabilization of the open-burst state that results from mutation of the possible catalytic base^{4,17}, Glu 1371 (a glutamate in the NBD2-head 'Walker B' motif), to Gln (E1371Q; Fig. 1b). Because CFTR channels do not open without ATP (see below), current decay upon the removal of ATP reflects channel closing, and its time course measures the open burst duration; closing was complete within about 1 s of ATP withdrawal for wild-type (WT) CFTR channels (mean open burst duration was less than 0.5 s), but had a time constant of 411 ± 64 s (mean $\pm$ s.e.m., $n = 16$) for mutant E1371Q channels (Fig. 1b). Correspondingly, in isolated NBD subunits impairment of ATP hydrolysis by the homologous Glu-to-Gln mutation induces the formation of very stable ATP-bound homodimers^{4,5,11,12}.

To test the dynamic domain rearrangements underlying the model (Fig. 1a), we investigated the interaction between residues on opposite sides of the predicted NBD1–NBD2 heterodimer interface. In crystals of ATP-bound prokaryotic NBD homodimers (MJ0796, the hyperthermophilic archaeon homologue of LolD, part of a putative lipoprotein transporter⁵, and MalK, the NBD of

the maltose transporter of *Escherichia coli*¹⁰) a hydrogen bond connects the residues corresponding to CFTR's Arg 555 (three positions after the signature sequence: LSGGQRR) in the NBD1 tail, and Thr 1246 (within the 'Walker A' phosphate-binding loop, GRTGSGKS) in the NBD2 head. We chose to target this pair of residues on the basis of a statistical analysis (Fig. 2b) of more than 10,000 NBD sequences, which suggests that positions corresponding to CFTR's Arg 555 and Thr 1246 are functionally coupled¹⁸. Partitioning the total multiple sequence alignment into subsets, on the basis of the side chain present at the site corresponding to Arg 555, yields one major subset with arginine, and another with lysine, at that position, both of which are potential hydrogen bond donors. The distribution of potential acceptor side chains at the position corresponding to Thr 1246 differs in these two subsets, the shorter lysine donor being more frequently paired with the longer asparagine acceptor, and the longer arginine donor being more

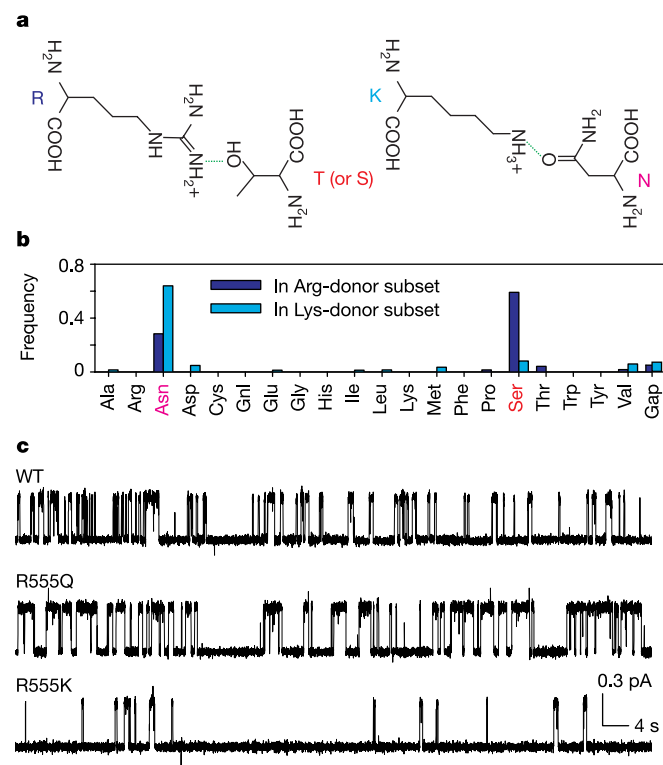


Figure 2 Statistical coupling analysis and electrophysiological recordings position CFTR's Arg 555 in the composite NBD2 catalytic site. **a**, Representation of hydrogen-bonded arginine-threonine (R–T) and lysine-asparagine (K–N) pairs. In serine and threonine (the side chain present in CFTR) the acceptor oxygen atom is positioned at the same distance from the peptide backbone. **b**, Amino acid frequencies at the 'head' position, equivalent to Thr 1246, in alignment subsets having Arg (blue bars) or Lys (cyan bars) at the 'tail' position. The total alignment contained 10,194 sequences (<http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00005>). **c**, Mutations at Arg 555 in NBD1 'tail' affect mean open burst (R555Q) and closed interburst (R555K) dwell times. R555Q opened with rates comparable to WT (R555Q, $\tau_{ib} = 2.84 \pm 0.53$ s (15)) despite expected loss of the interfacial hydrogen bond, suggesting possible compensation by other factors, for example the removal of repulsive electrostatic forces.

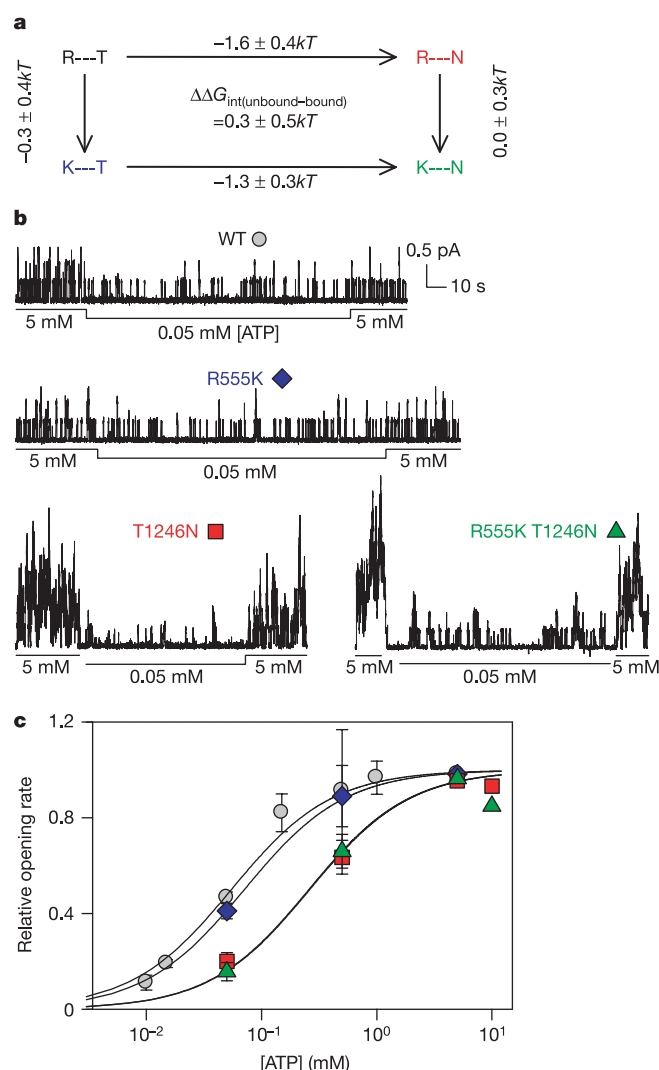


Figure 3 Arg 555 and Thr 1246 are not energetically coupled in channel closed states. **a**, Thermodynamic mutant cycle (each corner defined by the side chains at position 555 and 1246) using changes in apparent dissociation constant ($K_{0.5}$) as an estimate of the equilibrium constant for the reaction $C_2 \rightleftharpoons C_1 + \text{ATP}$ (Fig. 1a). Values show changes (means $\pm$ s.d.) in free energy difference driving this reaction under standard conditions, $\Delta G^{(C_1+ATP)} - C_2 = -kT \ln K_{0.5}$ (k is Boltzmann's constant, T is absolute temperature). **b**, Representative records from WT, single mutants R555K and T1246N, and double mutant R555K T1246N. Exposure to the test [ATP] (here 50 μ M) was bracketed by exposures to 5 mM ATP. **c**, Mean relative¹⁴ opening rates ($\pm$ s.e.m., with $2 \leq n \leq 25$).

frequently paired with the shorter serine acceptor (or the equivalent threonine, as found in CFTR) (Fig. 2b). This striking covariance suggests that the two sites have been subject to evolutionary pressure as a pair, rather than individually, so as to retain the ability to form a hydrogen bond that precisely distances the two α -carbons (Fig. 2a) in the dimer structure.

We first mutated Arg 555, and found that the charge-removing mutation R555Q slowed channel closing (mean open burst duration was 3.20 ± 0.35 s ($n = 18$) for R555Q, compared with 0.43 ± 0.02 s ($n = 32$) for WT; Fig. 2c), consistent with Arg 555 being part of the composite NBD2 catalytic site, where the ATP hydrolysis that times closing of WT channels occurs. Possibly, the positive charge of Arg 555 normally helps to stabilize the partial negative charges developed on the β - and/or γ -phosphate in the transition state for that ATP hydrolysis¹⁹; an equivalent mutation impairs ATP hydrolysis in the ABC multidrug transporter P-glycoprotein²⁰. Accordingly, the charge-conserving mutation R555K did not affect open burst duration (mean 0.39 ± 0.04 s ($n = 26$)). However, it substantially prolonged closed interburst duration (inversely related to opening rate) from 2.29 ± 0.46 s ($n = 16$) for WT to 8.53 ± 1.23 s ($n = 15$) for R555K (Fig. 2c, all measured at saturating [ATP]). On the basis of the evolutionary evidence suggesting the involvement of the side chain at this position in a conserved interaction (Fig. 2b), this slowing of channel opening could be explained if the R555K mutation were to weaken or remove a hydrogen bond between NBD1 and NBD2 that is absent

in the closed, ground, state but present in the transition state for the channel opening reaction; the resulting destabilization of the transition state would increase the activation free energy ($\Delta G^\ddagger$) for channel opening and hence decrease the opening rate.

To quantify this suspected interaction between Arg 555 and Thr 1246 side chains (Fig. 2a), we applied double mutant-cycle analysis^{6,21} (see Supplementary Information), after mutating Arg 555 to Lys and Thr 1246 to Asn, both individually and jointly. The WT protein, the two single mutants and the double mutant form the corners of a thermodynamic cycle (Figs 3a and 4a, d). If the two residues do not interact, the effects of mutating Arg 555 to Lys should be the same in a Thr 1246 background as in a T1246N background (and vice versa); that is, the effects of the single mutations should be independent and hence additive, and mutation-linked changes on parallel sides of the cycles should thus be equal. Any difference signifies, and quantifies, energetic coupling ($\Delta\Delta G_{\text{int}}$) between the two residues. Because changes in any path-independent variable can be used to evaluate the effects of the mutations, we used different kinetic measurements to assess energetic coupling between residues at positions 555 and 1246 at different stages of the channel gating cycle.

We first examined coupling in closed channels. The opening rate of WT CFTR shows simple Michaelis-Menten dependence on [ATP] (Fig. 3c), most probably reflecting ATP interaction with the NBD2-head binding site¹⁵. The measured maximal opening rate is relatively slow (0.27 s^{-1} , for pre-phosphorylated channels¹⁴; step

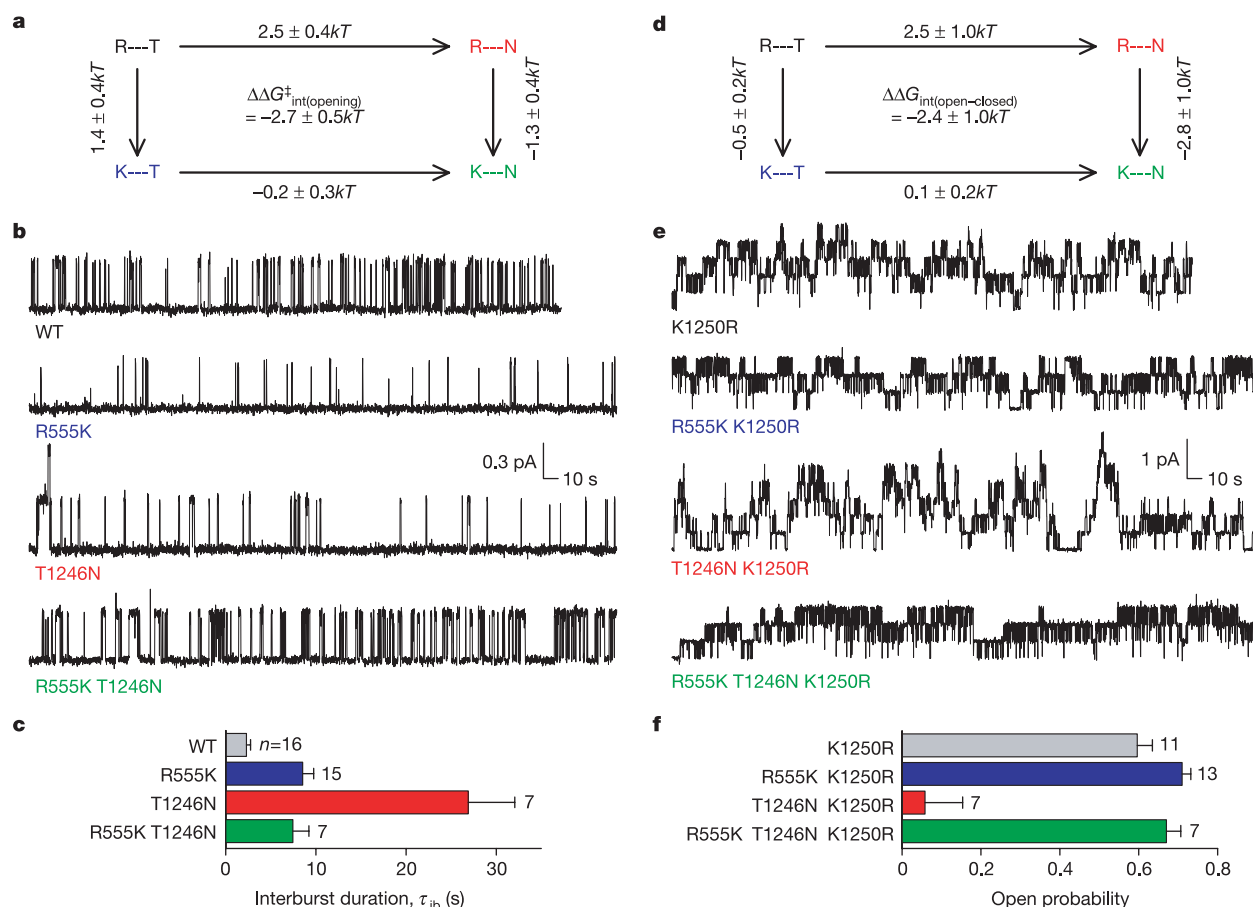


Figure 4 Energetic coupling between Arg 555 and Thr 1246 accompanies channel opening. **a**, Thermodynamic cycle showing changes (means $\pm$ s.d.) in the activation energy barrier for opening ($\Delta\Delta G^\ddagger$). **b**, Representative records. **c**, Mean closed interburst duration ($\pm$ s.e.m.). **d**, Thermodynamic cycle showing changes (means $\pm$ s.d.) in stability

of open state with respect to the closed state, $\Delta\Delta G_{\text{open-closed}}$, calculated from P_o . **e**, Representative records. Current levels of the triple mutant R555K T1246N K1250R did not change when [ATP] was increased to 10 mM, indicating that 5 mM [ATP] was saturating. **f**, Mean P_o ($\pm$ s.e.m.).

C₂ to O in Fig. 1a), allowing the preceding ATP binding step (C₁ to C₂) to reach a steady state not far from equilibrium. Consequently, the apparent affinity for ATP ($K_{0.5}$) provides a good estimate of its dissociation constant at the NBD2 site on the closed channel (C₂). Changes in $K_{0.5}$ could therefore be used to quantify the effects of mutations and so assess coupling between Arg 555 and Thr 1246 in closed channels (Fig. 3). The apparent affinity for ATP was little influenced by the mutation R555K (R555K $K_{0.5} = 71 \pm 14 \mu\text{M}$ versus WT $K_{0.5} = 55 \pm 5 \mu\text{M}$), but was reduced by the mutation T1246N (T1246N $K_{0.5} = 261 \pm 49 \mu\text{M}$) by the same extent in the WT background as in the R555K background (R555K T1246N $K_{0.5} = 257 \pm 51 \mu\text{M}$). The closely similar effects of mutations on parallel sides of the double mutant cycle signify a coupling energy not significantly different from zero ($\Delta\Delta G_{\text{int(unbound-bound)}} = 0.31 \pm 0.55 kT$). So, either Arg 555 and Thr 1246 do not interact when CFTR channels are closed, or they do interact but with identical coupling energies before, and after, binding of the second ATP (that is in state C₁ and in state C₂). But, given that the residue corresponding to Thr 1246 is observed to directly contact the γ -phosphate oxygens in all ATP-bound crystal structures, both monomeric and dimeric^{3,5,10,22,23}, it is highly improbable that there is significant energetic coupling between Arg 555 and Thr 1246 that is unaffected by ATP binding. We therefore conclude that the two target side chains do not interact in the closed-channel conformations (either with or without bound ATP), and hence that ATP binding occurs before the formation of a closely apposed NBD1–NBD2 dimer. Indeed, in dimeric crystals the bound ATP molecules are buried within the interface^{5,10}, implying that access to the binding sites must occur in a different conformation.

We next tested for energetic coupling as CFTR channels approach the open burst state, by determining changes in activation free energies for channel opening (in the presence of saturating [ATP]). The slowing of opening caused by the R555K mutation (Fig. 2c, above) corresponded to a $1.4 \pm 0.4 kT$ increase in the activation energy barrier. The T1246N mutation also greatly slowed channel opening, increasing the energy barrier by $2.5 \pm 0.4 kT$. However, fast opening was partly restored when the two mutations were introduced simultaneously (Fig. 4b, c). Comparing changes in activation-energy barrier height caused by each mutation in the WT, and in the mutant background, we obtain an energetic coupling between residues at positions 555 and 1246 of $\Delta\Delta G_{\text{int(opening)}}^{\ddagger}$ of $-2.7 \pm 0.5 kT$. The negative sign is consistent with Arg 555 and Thr 1246 in WT CFTR forming a stabilizing interaction, present in the transition state but not in the closed, ground, state (the above conclusion that these two residues do not interact in closed channels allows us to rule out the alternative interpretation, that the negative coupling energy could arise from a destabilizing interaction in the closed state that is lost in the transition state for channel opening; see also Supplementary Information). The presence of the postulated hydrogen bond between Arg 555 and Thr 1246 in the transition state could represent that stabilizing interaction: either mutation alone would remove the hydrogen bond and destabilize the transition state, whereas both mutations together would restore favourable geometry for hydrogen bond formation (Fig. 2a) and so speed up channel opening.

Interaction between Arg 555 and Thr 1246 in the open burst state, rather than approaching it, was monitored by measuring channel open probability (P_o) to estimate the relative stability of the closed and open states, after reducing the channel gating scheme to a simple closed–open equilibrium. For the scheme in Fig. 1a, the latter condition holds, at saturating [ATP], if channel closing from open bursts by means of the hydrolytic pathway (O to C₁) is precluded. In other ABC-ATPases, mutating the key lysine in the phosphate-binding loop to arginine drastically reduces or abolishes hydrolysis (see, for example, ref. 24) and, as would be predicted if hydrolysis at CFTR's NBD2 catalytic site were markedly slowed, CFTR channels carrying the corresponding mutation (K1250R)

have prolonged open burst durations (Fig. 4e; mean time constant of current decay upon ATP removal $\tau = 9.3 \pm 0.5 \text{ s}$; $n = 49$). Introducing the T1246N mutation into the K1250R background decreased P_o , corresponding to destabilization of the open burst state by $2.5 \pm 1.0 kT$ with respect to the closed state. However, adding the R555K mutation to T1246N–K1250R channels restored high stability of the open state (Fig. 4e, f). The coupling energy obtained from the mutant cycle (Fig. 4d) was again negative in sign ($\Delta\Delta G_{\text{int(open-closed)}} = -2.4 \pm 1.0 kT$), which is consistent with the presence of a stabilizing interaction (for example a hydrogen bond) between Arg 555 and Thr 1246 in the open burst state that is absent from the closed state.

Thus, as conformational changes gate CFTR's transmembrane Cl[−]-ion permeation pathway from closed to open, two residues, on opposite sides of the anticipated NBD1–NBD2 heterodimer interface, go from being independent to being energetically coupled, most probably by forming a hydrogen bond. The changes in the relative positions of CFTR's NBD1 and NBD2 implied by our results need not be large. They need be no larger than the relatively small differences observed between members of NBD pairs, juxtaposed in a 'head-to-tail' fashion, in crystal structures of nucleotide-bound versus nucleotide-free conformations. The residues corresponding to Arg 555 and Thr 1246 are close enough (2.7 Å) to form a hydrogen bond in the ATP-bound crystals of LolD (MJ0796; ref. 5) and MalK (PDB ID no. 1Q12; ref. 10) and are found only a few ångströms further apart in nucleotide-free structures (for example, 5.7 Å in BtuCD (ref. 9); 8.5 Å in MalK, PDB ID no. 1Q1B (ref. 10)), in which the NBDs share a limited contact surface but are kept in a dimeric arrangement by interactions occurring through contiguous domains (see also ref. 13). Nevertheless, however small the local change at CFTR's NBD2 catalytic site, it is large enough to cause transmission of the long-range signal that results in opening of the channel pore.

By clearly linking NBD dimerization state to the disposition of the transmembrane domains in an intact, functioning, human ABC protein, our results establish dynamic NBD dimerization as the molecular mechanism that couples ATP binding and hydrolysis cycles to cyclic changes in the transmembrane domains. Conservation of the structural underpinnings of the hydrogen bond highlighted here implies that NBD dimerization occurs during the duty cycle of most, if not all, ABC proteins, although with subfamily-specific consequences. Thus, the tightly dimerized NBD conformation corresponds to the open-burst state in CFTR channels (Fig. 1), but in exporters or importers to conformations of the membrane domains that release drugs to the exterior²⁵, or receive substrate from the substrate-binding protein²⁶, respectively, and in DNA-repair ABC ATPases to relevant conformations in the DNA-binding regions²⁷. □

Methods

Oocyte expression system and experimental set-up

Human epithelial CFTR was mutated in an oocyte expression vector (pGEMHE-WT¹⁴) using QuikChange (Stratagene), and DNA templates were linearized with *NheI* before RNA transcription *in vitro* (mMessage mMach; Ambion). Inside-out patches were excised from *Xenopus laevis* oocytes expressing WT or mutant CFTR channels and currents were recorded as described¹⁴. Bath (cytosolic) solution contained (in mM): 138 *N*-methyl-D-glucamine (NMDG), 2 MgCl₂, 5 HEPES, 0.5 EGTA, 136 sulphamic acid (pH 7.1). Pipette (extracellular) solution contained (in mM): 138 NMDG, 2 MgCl₂, 5 HEPES, 136 HCl (pH 7.4). Membrane potential was held at +50 mV (−50 mV pipette potential), and outward currents were filtered at 200 Hz and digitized at 1 kHz. Bath solution flowed continuously (about 0.5 ml min^{−1}) and solution was exchanged by using computer-driven solenoid valves. Channels were activated by exposure to 5 mM ATP, magnesium salt, in the presence of 300 nM purified cAMP-dependent protein kinase catalytic subunit.

Data analysis

Digitized records were baseline-subtracted and idealized by half-amplitude threshold crossing; the events list was used to extract mean open burst (τ_o) and closed interburst (τ_{ib}) dwell times by simultaneous maximum-likelihood fitting to dwell-time histograms at all conductance levels²⁸. Openings of CFTR channels are 'open bursts' occasionally

interrupted by short-lived (non-conducting) 'flickery' closures, with low frequency and brief duration independent of ATP binding and hydrolysis. We refer to the 'open burst' as a single, composite state: channels 'open' upon entering a burst and 'close' upon entering the long-lived, interburst, closed state. However, the maximum-likelihood fit analysis yields estimates for all parameters, including flicker duration (τ_F) and number of flickers per burst (n_F) (see Supplementary Table).

An artificial dead time of 4.5 ms was imposed to implement a correction for events missed because of limited bandwidth²⁸. To obtain τ_{ib} measurements, only patches containing at most two simultaneously open channels were used. For constructs with very low P_o (R555K and T1246N), we could not exclude the presence of unseen channels in the patch (even though the records lasted on average 6–7 min). The prolonged τ_{ib} values we extract for R555K and T1246N channels are therefore most probably underestimates, and the real effects of the mutations are more severe (and, hence, $|\Delta\Delta G_{int(opening)}^\ddagger|$ is actually larger) than the values we report.

P_o values (Fig. 4e, f) were estimated, on the assumption that channels are identical and independent and that the number of open channels in the patch is a random variable with binomial distribution, from the relationship between mean current ($I = NP_o i$, where I is the mean current, N is the number of channels present in the patch, and i is the single-channel current) and variance of the current ($\sigma^2(I) = Ni^2 P_o(1 - P_o)$, where $\sigma^2(I)$ is the variance of the current).

Relative opening rates (Fig. 3c) were obtained as described¹⁴. In brief, the maximum-likelihood fit was performed on the assumption that the number of channels present in a given patch (N) was equivalent to the maximum number of simultaneously open channels observed, during all test and reference conditions, in that patch. The opening rate obtained from the segments during exposure to the test [ATP] was then normalized to that obtained during the bracketing exposures to 5 mM ATP, giving a relative value that was little sensitive to errors in our estimate of N .

Received 11 October; accepted 22 December 2004; doi:10.1038/nature03313.

- Gunderson, K. L. & Kopito, R. R. Conformational states of CFTR associated with channel gating: the role of ATP binding and hydrolysis. *Cell* **82**, 231–239 (1995).
- Carson, M. R., Travis, S. M. & Welsh, M. J. The two nucleotide-binding domains of cystic fibrosis transmembrane conductance regulator (CFTR) have distinct functions in controlling channel activity. *J. Biol. Chem.* **270**, 17111–17117 (1995).
- Hopfner, K. P. *et al.* Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. *Cell* **101**, 789–800 (2000).
- Moody, J. E., Millen, L., Binns, D., Hunt, J. F. & Thomas, P. J. Cooperative, ATP-dependent association of the nucleotide binding cassettes during the catalytic cycle of ATP-binding cassette transporters. *J. Biol. Chem.* **277**, 21111–21114 (2002).
- Smith, P. C. *et al.* ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. *Mol. Cell* **10**, 139–149 (2002).
- Serrano, L., Horovitz, A., Avron, B., Bycroft, M. & Fersht, A. R. Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry* **29**, 9343–9352 (1990).
- Davidson, A. L. & Chen, J. ATP-binding cassette transporters in bacteria. *Annu. Rev. Biochem.* **73**, 241–268 (2004).
- Higgins, C. F. & Linton, K. J. The ATP switch model for ABC transporters. *Nature Struct. Mol. Biol.* **11**, 918–926 (2004).
- Locher, K. P., Lee, A. T. & Rees, D. C. The *E. coli* BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science* **296**, 1091–1098 (2002).
- Chen, J., Lu, G., Lin, J., Davidson, A. L. & Quirocho, F. A. A tweezers-like motion of the ATP-binding cassette dimer in an ABC transport cycle. *Mol. Cell* **12**, 651–661 (2003).
- Janas, E. *et al.* The ATP hydrolysis cycle of the nucleotide-binding domain of the mitochondrial ATP-binding cassette transporter Mdl1p. *J. Biol. Chem.* **278**, 26862–26869 (2003).
- Verdon, G. *et al.* Formation of the productive ATP-Mg²⁺-bound dimer of GlcV, an ABC-ATPase from *Sulfolobus solfataricus*. *J. Mol. Biol.* **334**, 255–267 (2003).
- Horn, C., Bremer, E. & Schmitt, L. Nucleotide dependent monomer/dimer equilibrium of OpuAA, the nucleotide-binding protein of the osmotically regulated ABC transporter OpuA from *Bacillus subtilis*. *J. Mol. Biol.* **334**, 403–419 (2003).
- Vergani, P., Nairn, A. C. & Gadsby, D. C. On the mechanism of MgATP-dependent gating of CFTR Cl[−] channels. *J. Gen. Physiol.* **121**, 17–36 (2003).
- Basso, C., Vergani, P., Nairn, A. C. & Gadsby, D. C. Prolonged nonhydrolytic interaction of nucleotide with CFTR's NH₂-terminal nucleotide binding domain and its role in channel gating. *J. Gen. Physiol.* **122**, 333–348 (2003).
- Aleksandrov, L., Aleksandrov, A. A., Chang, X. B. & Riordan, J. R. The first nucleotide binding domain of cystic fibrosis transmembrane conductance regulator is a site of stable nucleotide interaction, whereas the second is a site of rapid turnover. *J. Biol. Chem.* **277**, 15419–15425 (2002).
- Tomblin, G., Bartholomew, L. A., Urbatsch, I. L. & Senior, A. E. Combined mutation of catalytic glutamate residues in the two nucleotide binding domains of P-glycoprotein generates a conformation that binds ATP and ADP tightly. *J. Biol. Chem.* **279**, 31212–31220 (2004).
- Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).
- Knowles, J. Enzyme-catalyzed phosphoryl transfer reactions. *Annu. Rev. Biochem.* **49**, 877–919 (1980).
- Bakos, É. *et al.* Characterization of the human multidrug resistance protein containing mutations in the ATP-binding cassette signature region. *Biochem. J.* **323**, 777–783 (1997).
- Fersht, A. *Structure and Mechanism in Protein Science* (W. H. Freeman, New York, 1999).
- Hung, L. W. *et al.* Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature* **396**, 703–707 (1998).
- Verdon, G., Albers, S. V., Dijkstra, B. W., Driessen, A. J. M. & Thunnissen, A.-M. W. Crystal structures of the ATPase subunit of the glucose ABC transporter from *Sulfolobus solfataricus*: nucleotide-free and nucleotide-bound conformations. *J. Mol. Biol.* **330**, 343–358 (2003).
- Lerner-Marmarosh, N., Gimi, K., Urbatsch, I. L., Gros, P. & Senior, A. E. Large scale purification of detergent-soluble P-glycoprotein from *Pichia pastoris* cells and characterization of nucleotide binding

properties of wild-type, Walker A, and Walker B mutant proteins. *J. Biol. Chem.* **274**, 34711–34718 (1999).

- Qian, Y.-M. *et al.* Characterization of binding of leukotriene C₄ by human multidrug resistance protein 1. Evidence of differential interactions with NH₂- and COOH-proximal halves of the protein. *J. Biol. Chem.* **276**, 38636–38644 (2001).
- Austermuhle, M. L., Hall, J. A., Klug, C. S. & Davidson, A. L. Maltose-binding protein is open in the catalytic transition state for ATP hydrolysis during maltose transport. *J. Biol. Chem.* **279**, 28243–28250 (2004).
- Hopfner, K.-P. & Tainer, J. A. Rad50/SMC proteins and ABC transporters: unifying concepts from high-resolution structures. *Curr. Opin. Struct. Biol.* **13**, 249–255 (2003).
- Csanády, L. Rapid kinetic analysis of multichannel records by a simultaneous fit to all dwell-time histograms. *Biophys. J.* **78**, 785–799 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Csanády and G. Szakács for discussion. The work was supported by an NIH grant to D.C.G.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.C.G. (gadsby@rockefeller.edu) or to P.V. (paola.vergani@rockefeller.edu).

Force generation by mammalian hair bundles supports a role in cochlear amplification

H. J. Kennedy^{1,2}, A. C. Crawford³ & R. Fettiplace¹

¹Department of Physiology, University of Wisconsin Medical School, Madison, Wisconsin 53706, USA

²Department of Physiology, University of Bristol, Bristol BS8 1TD, UK

³Department of Physiology, Cambridge University, Cambridge CB2 3EG, UK

It is generally accepted that the acute sensitivity and frequency discrimination of mammalian hearing requires active mechanical amplification of the sound stimulus within the cochlea¹. The prevailing hypothesis is that this amplification stems from somatic electromotility of the outer hair cells attributable to the motor protein prestin^{2,3}. Thus outer hair cells contract and elongate in synchrony with the sound-evoked receptor potential^{4,5}. But problems arise with this mechanism at high frequencies, where the periodic component of the receptor potential will be attenuated by the membrane time constant. On the basis of work in non-mammalian vertebrates, force generation by the hair bundles has been proposed as an alternative means of boosting the mechanical stimulus^{6,7}. Here we show that hair bundles of mammalian outer hair cells can also produce force on a submillisecond timescale linked to adaptation of the mechanotransducer channels. Because the bundle motor may ultimately be limited by the deactivation rate of the channels, it could theoretically operate at high frequencies. Our results show the existence of another force generator in outer hair cells that may participate in cochlear amplification.

The mammalian hearing organ, the cochlea, contains two classes of sensory receptor, namely inner and outer hair cells, with disparate functions. Acoustic information is relayed primarily via the inner hair cells and their synapses on the auditory nerve afferents, whereas outer hair cells (OHCs) act in parallel to boost the stimulus by electromechanical feedback¹. The locus of hair cell transduction is the hair bundle⁸, which in OHCs comprises three ranks of modified microvilli, known as stereocilia, in a 'V'-shaped array. Deflection of the bundle towards the point of the 'V' opens mechanoelectrical transducer (MET) channels near the tips of the stereocilia,

permitting influx of K^+ and Ca^{2+} to evoke a depolarizing receptor potential^{8–11}. The MET channels open rapidly but then adapt with a submillisecond time constant¹⁰ similar to that reported in non-mammalian vertebrates¹². In the turtle, fast adaptation reflects calcium-dependent re-closure of the MET channels, which can in turn elicit a mechanical reaction that moves the hair bundle¹³. However, there is little evidence for a mechanical correlate of adaptation in mammals.

The mechanical properties of OHC hair bundles in isolated cochleas of neonatal rats were assayed by applying force stimuli with a calibrated flexible glass fibre (Fig. 1) and measuring the ensuing displacement of its image on a photodiode pair^{14,15}. Excitatory stimuli elicited an MET current with fast onset and adaptation, the peak amplitude of the current showing sigmoidal dependence on bundle displacement. Surprisingly, the force–displacement relationship of the hair bundle did not behave like that of a simple spring, but was very nonlinear and became increasingly compliant over the range where the MET channels were gated (Fig. 2). Furthermore, the nonlinearity developed with a time course similar to fast adaptation of the MET current. The slow offsets for the three largest steps (Fig. 2a) had an electrical correlate in the slow rebound of the MET current, reflecting recovery from adaptation. In contrast, the displacement of the flexible fibre when not attached to the bundle was fast, and its amplitude was proportional to the applied force for movements less than $0.8 \mu\text{m}$ (Fig. 1). Nonlinear hair bundle mechanics were observed in 12 out of 15 cells studied. This type of behaviour was first reported in non-mammalian vertebrates where it was attributed to the ‘gating compliance’, a decrease in stiffness associated with opening of the MET channels^{15,16}. Nonlinearities in hair bundle mechanics, probably of similar origin, have also been seen in the mammalian cochlea but were not linked to adaptation¹⁷.

The results were analysed in terms of the gating-spring model¹⁵, which predicts a relationship between the applied force (F_B) and bundle displacement (X) given by:

$$F_B = XK_s - Ap_o(X) + F_o \quad (1)$$

where K_s is the passive linear stiffness, p_o is the probability of opening of the MET channels, and A and F_o are constants. The negative term in equation (1) signifies an active component in which channel gating generates a force in the same direction as the

imposed displacement, causing hair bundle stiffness to decrease with channel opening to reach a minimum when p_o is ~ 0.5 (ref. 15). Fits of equation (1) to the results deviated from the gating-spring model in that the constant A increased as adaptation progressed, a discrepancy evident in the displacement dependence of bundle stiffness (Fig. 2d). A further difference is that A was much bigger than predicted by the gating-spring model. In that model, A is the product of the number of MET channels and the single-channel gating force, which is less than 30 pN for OHCs (ref. 18), whereas values 20- to 100-fold larger were needed for the fits in Fig. 2c. Nevertheless, our results are still consistent with force production being linked to the probability of opening of the MET channels.

In some OHCs, the force–displacement relationship possessed a negative slope region (Fig. 2d) similar to that seen in frog hair cells¹⁹. However, in the frog the nonlinearity was associated with channel activation and was effectively instantaneous; here the nonlinearity was time dependent and developed with adaptation. In five cells, more extreme behaviour was observed: for a range of stimuli the displacement of the end of the flexible fibre attached to the hair bundle was larger than that of the end cemented to the piezoelectric device (Fig. 3). Thus the force–displacement relationship became negative as adaptation progressed, indicating that the hair bundle was doing work on the fibre. Maximum force generation, estimated as the difference between steady state and instantaneous force–displacement plots at fixed displacement, was 517 ± 96 pN in five cells.

According to the analysis, force generation by the bundle reflects events at the level of the MET channel. Consistent with this notion, experimental manipulations affecting channel gating altered the force–displacement relationship. One manipulation is lowering extracellular Ca^{2+} , which is known to slow and reduce adaptation in auditory hair cells of both turtle²⁰ and rat¹⁰. Similarly, lowering Ca^{2+} reversibly reduced the hair bundle mechanical nonlinearity and slowed its onset (Fig. 4). Altering the external Ca^{2+} concentration from 1.5 to 0.02 mM shifted the $p_o(X)$ relationship in the

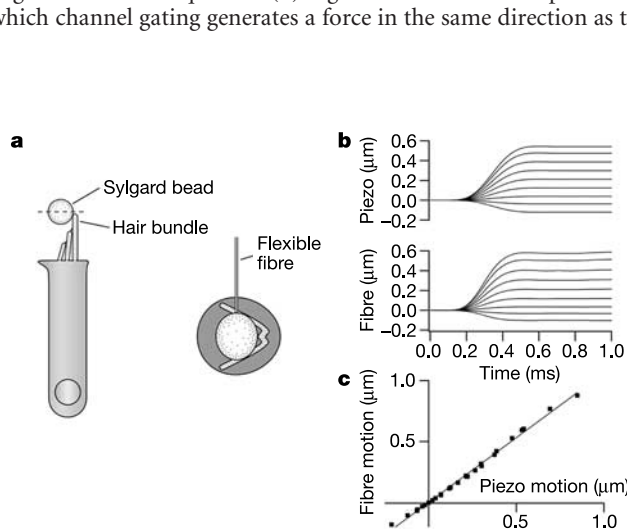


Figure 1 Method of hair bundle stimulation. **a**, Relationship between the Sylgard bead and the $5 \mu\text{m}$ OHC bundle, viewed from the side and from the top of the cell. **b**, Time course of stimulus onset: top, the driving voltage to the piezoelectric stack, shaped with an eight-pole Bessel filter at 1.5 kHz; bottom, the resulting motion of the fibre tip when not attached to the bundle. The fibre motion had a 10–90% rise time of 0.2 ms, and was not slowed by viscous drag. **c**, Linearity of fibre motion with the amplitude of the piezoelectric drive. Slope of fitted line, 1.06 ± 0.02 (1 s.d.).

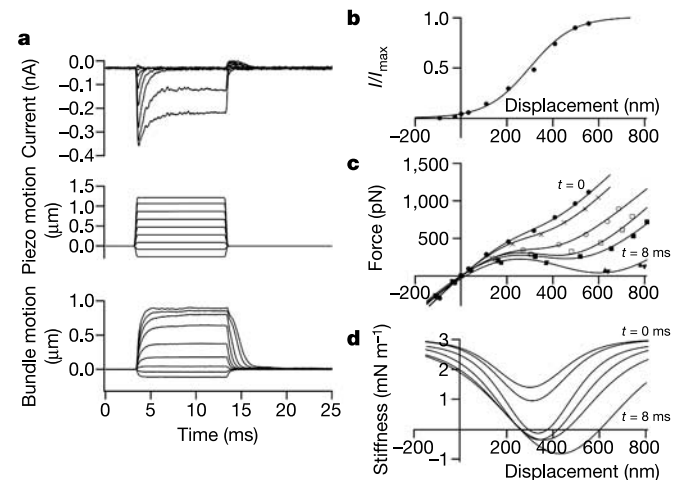


Figure 2 Mechanical properties of the OHC hair bundle. **a**, Top; MET currents in a P11 rat for stimulation with flexible fibre. Also shown are the movements of the fibre end attached to the piezo (middle) and of the hair bundle (bottom). **b**, Peak MET current (I) versus displacement (X) (data points) fitted with a Boltzmann relation: $I/I_{\text{max}} = 1/(1 + \exp(-(X - X_o)/X_o))$, where $I_{\text{max}} = 0.37$ nA, $X_o = 300$ nm and $X_e = 100$ nm. **c**, Force–displacement plots at successive times (t) after peak current: $t = 0$ (filled circles), 0.07, 0.27, 0.47, 0.67, 3.9, 8 ms (filled triangles). Theoretical fits with equation (1) using $p_o(X)$ = the Boltzmann relation from **b** and $K_s = 3$ mN m⁻¹. **d**, Slope stiffness of the bundle from differentiating fits in **c**.

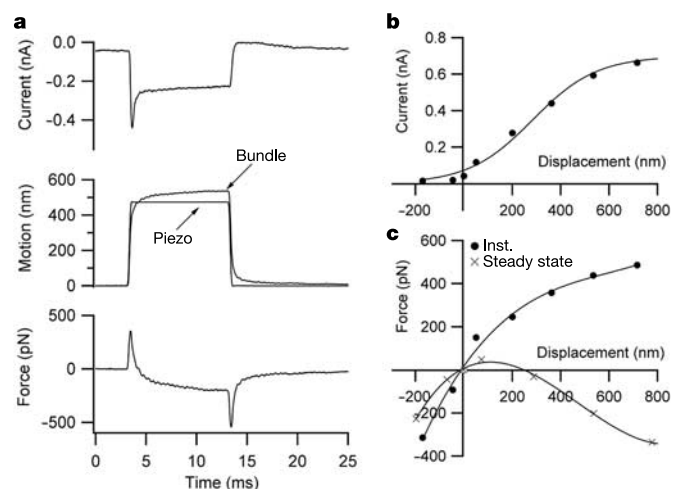


Figure 3 Force generation by the OHC bundle. **a**, Top: MET current for flexible fibre stimulation. Middle, superimposed records of the motion of the end of the fibre connected to the piezo and of the end of fibre attached to the bundle shows that the bundle moves more than the piezo, indicating extra force generation by the bundle. Bottom, the force experienced by the fibre, calculated from the difference in the piezo and bundle displacements times the fibre stiffness (2.8 mN m^{-1}). Ripples reflect high-frequency resonance in the piezo. **b**, MET current versus displacement fitted with a Boltzmann relation with $I_{\text{max}} = 0.67 \text{ nA}$, $X_0 = 284 \text{ nm}$ and $X_\infty = 132 \text{ nm}$. **c**, Force-displacement plots at peak inward current, $t = 0$ (inst.), and at $t = 8 \text{ ms}$ (steady-state).

negative direction by 200 nm, reduced the force constant A by 21% and doubled the passive stiffness K_s (equation (1)). Force-displacement relationships in high Ca^{2+} (Figs 2c and 4b) and low Ca^{2+} (Fig. 4a) were used to determine the time course of the mechanical effect. The difference in force between the instantaneous relation (time $t = 0$, measured at the peak current) and the plots at subsequent times were measured at a fixed displacement corresponding approximately to the point of maximum compliance. The

change in force was plotted against time (Fig. 4d), and fitted with time constants of 0.3 ms (1.5 mM Ca^{2+}) and 0.6 ms (0.02 mM Ca^{2+}). (In low Ca^{2+} there was also a small slower component). These values agreed well with the main component of fast adaptation: 0.28 ms in 1.5 mM Ca^{2+} and 0.68 ms in 0.02 mM Ca^{2+} . Similar agreement was obtained in three other hair cells, from which we conclude that bundle force develops with the same time course as fast adaptation. Lowering Ca^{2+} also increased passive hair bundle stiffness (K_s in equation (1)), inferred from instantaneous plots for negative displacements. In four cells, a mean stiffness of $1.9 \pm 0.6 \text{ mN m}^{-1}$ in 1.5 mM Ca^{2+} was increased to $5.5 \pm 1.0 \text{ mN m}^{-1}$ in 0.02 mM Ca^{2+} . The passive stiffness averaged over all the cells was $3.6 \pm 0.2 \text{ mN m}^{-1}$ ($n = 14$; 1.5 mM Ca^{2+}), similar to previous measurements on mammalian OHCs^{17,21–23}. Ca^{2+} could affect the stiffness of either the stereociliary pivot or the gating spring.

Our new observations on the magnitude ($>500 \text{ pN}$) and speed ($<1 \text{ ms}$) of force generation by hair bundles of cochlear OHCs raise several questions. Principally, what mechanism intrinsic to the bundle is capable of producing large forces with such fast kinetics? A possible candidate is one of the unconventional myosins found in the hair bundle, myosin XVa, localized to the tips of the stereocilia²⁴, or myosin VIIa that is reported to affect OHC adaptation²⁵. However, force production by any myosin is unlikely to occur sufficiently fast to explain our results. An alternative is the MET channel itself, which during gating must undergo conformational changes on a microsecond timescale *in vivo* to transduce the high frequencies encoded by the mammalian cochlea (up to 60 kHz in the rat²⁶). Although force production by OHC bundles of rats, like turtles, may involve MET channel adaptation, the underlying mechanism must be different in the two cases because the movements are of different polarity. In the turtle, fast adaptation is associated with recoil of opposite direction to the imposed displacement¹³, whereas in the mammal the bundle moves further in the same direction. In this respect, active bundle motion in the mammal resembles that linked to slow adaptation in frogs²⁷, though the time course is 100-fold faster. In a specific mechanism, fast,

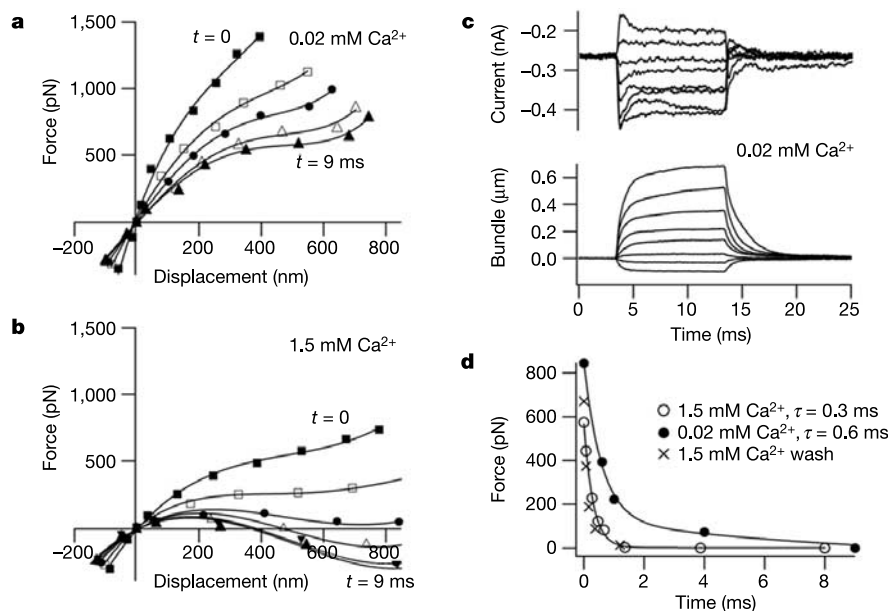


Figure 4 Effects of Ca^{2+} on OHC bundle mechanics. **a**, Force-displacement relationships in 0.02 mM Ca^{2+} for times after the peak inward current ($t = 0$) to the end of stimulus ($t = 9 \text{ ms}$). Figure 2c shows prior high Ca^{2+} controls. **b**, Return controls (wash) in 1.5 mM Ca^{2+} . **c**, MET currents and bundle movements in 0.02 mM Ca^{2+} . **d**, Time course of force production in 1.5 mM and 0.02 mM Ca^{2+} , measured from **a**, **b** and Fig. 2c

at displacements corresponding to minimum slope. To display like adaptation, all forces have been referred to that in the steady state. Fits have time constants (τ) of 0.3 ms (1.5 mM Ca^{2+} , prior control and wash) and 0.6 ms plus a small component of 9.8 ms (0.02 mM Ca^{2+}).

calcium- or stretch-activated, rocking of a molecule in series with the channel gate must relieve the applied force and cause adaptation. It is envisaged that this mechanical event is transmitted via the channel to reduce tension in the tip link, hence allowing further motion of the bundle.

A second question is how the hair bundle motor interacts with the somatic prestin motor. Because the somatic motor is controlled by changes in membrane potential, its role may be limited to low frequencies owing to filtering imposed by the cell time constant (though ways of circumventing this restriction have been proposed²⁸). Measured values of OHC time constants yield corner frequencies for the membrane filter of ~ 1 kHz (refs 29, 30). Our measurements were performed under voltage-clamp, precluding involvement of prestin: control of hair bundle mechanics must occur via activation and adaptation of the MET current. We have previously measured OHC adaptation time constants as fast as 120 μ s, but when corrected for the environmental conditions *in vivo* it may be less than 50 μ s at the apical position assayed¹⁰. Furthermore, if the adaptation time constant in the mammalian cochlea varies inversely with characteristic frequency, as it does in non-mammals¹³, it may be even faster at higher-frequency locations. However, better measuring techniques will be needed to demonstrate that force production by the hair bundle can follow at speeds of 50 μ s or less. The OHC bundle motor could supplement force generation by the somatic motor. More speculatively, it may constitute the principal mechanism underlying the cochlear amplifier, in which case the prestin motor may serve as a coarse adjustment, ensuring that the hair bundle is always optimally positioned to operate within its limited dynamic range. □

Methods

Experiment preparation and recording

These methods were identical to those described previously¹⁰. Rats, 7 to 11 days post-natal, were killed by cervical dislocation, decapitated and the temporal bone removed using procedures approved by the Animal Care Committee of the University of Wisconsin. The apical and middle turns of the cochlea were excised after unpeeling the stria vascularis, incubating in saline containing 50 μ g ml⁻¹ of proteinase (Sigma type XXIV), and removing the tectorial membrane. The preparation was viewed through a 40 $\times$ LWD water-immersion objective on a Zeiss Axioskop FS microscope, and perfused with saline of composition (in mM): 152 NaCl, 6 KCl, 1.5 CaCl₂, 2 Na-pyruvate, 10 glucose and 10 Na-HEPES, pH 7.4. The Ca²⁺ concentration was reduced to 20 μ M by buffering with 4 mM HEDTA. First or second row outer cells were voltage-clamped at -84 mV (temperature, 19–23 °C) with borosilicate patch electrodes connected to an Axopatch 200A amplifier. Patch pipettes were filled with a solution (in mM): 142 CsCl, 3.5 MgCl₂, 1 EGTA, 5 Na₂ATP, 0.5 Na₂GTP and 10 Cs-HEPES pH 7.2. Recordings were made at a location in the apical turn corresponding to a characteristic frequency of ~ 5 kHz (ref. 26). The recording time constant, with 70% series-resistance compensation, was between 40 and 80 μ s. All records are averages of 10 stimulus presentations, and measurements are given as mean $\pm$ 1 s.e.m.

Measurement of mechanical properties

The mechanical properties of hair bundles were measured with calibrated flexible glass fibres (~ 30 μ m in length; stiffness, 1–3 mN m⁻¹) as previously described^{13,14}. The proximal end of the flexible fibre was cemented through the side of a nylon screw attached to a piezoelectric stack actuator (Physik Instrumente), the fibre thus being perpendicular to the axis of motion of the actuator. The fibre was introduced along the axis of the cochlea, and its distal tip was coated with a ~ 3 - μ m-diameter bead of Sylgard (silicone elastomer; Dow-Corning Corporation) that fitted into the 'V'-shaped hair bundle of the OHC (Fig. 1). While applying small positive deflections, the bead was advanced onto the bundle until it just evoked a transducer current. The time course of motion of the fibre was monitored by imaging the edge of the Sylgard bead on a pair of photodiodes (Centronics LD 2–5) at 400 $\times$ total magnification. Owing to its higher refractive index, the Sylgard bead had greater contrast than the bundle, and therefore gave the dominant signal. Because the equator of the bead (dashed line in Fig. 1a), corresponding to the plane of focus, may have been slightly above the tips of the tallest stereocilia, the displacement of the bundle tip was over-estimated by measuring the motion of the bead. With a typical geometry shown in Fig. 1a, the displacement error was no more than 10%, but this error did not affect the calculation of applied force. The photodiodes were mounted on a piezoelectric bimorph, deflections of which were used to calibrate the photocurrent throughout the recording¹⁶.

For each experiment, motion of the fibre when not attached to the hair bundle was monitored to calibrate the frequency response and linearity of the stimulator and detector. With the voltage step to the piezoelectric device filtered at 1.5–2 kHz, displacement of the fibre tip in free solution above the bundle had a rise time of 150–200 μ s, very similar to that of the driving voltage (Fig. 1). The speed of the fibre was therefore limited largely by the

filtering applied to the piezoelectric stack and not by viscous drag on the fibre or bead. When the fibre was attached to the bundle, its drag time constant (equal to the viscous resistance divided by mechanical stiffness) was probably even smaller because of the larger combined stiffness of the fibre and bundle. To construct the force–displacement plots, the first reading of bundle motion was taken at a time (denoted by $t = 0$ in each figure) corresponding to the peak of the low-level MET current. The current reached a peak 300 μ s after the start of the response, which was approximately three times the equivalent time constant of the filter settings (80–106 μ s). Excluding the first 300 μ s of response ensured that viscous drag was not contributing to the measured time course of the bundle motion¹⁵. The small contribution of viscous drag is emphasized by the results in Fig. 4, where lowering Ca²⁺ increased hair bundle stiffness (which should have reduced the drag time constant), but slowed the bundle motion.

Received 3 November 2004; accepted 17 January 2005; doi:10.1038/nature03367.

Published online 6 February 2005.

- Dallos, P. The active cochlea. *J. Neurosci.* **12**, 4575–4585 (1992).
- Zheng, J. *et al.* Prestin is the motor protein of cochlear outer hair cells. *Nature* **405**, 149–155 (2000).
- Liberian, M. C. *et al.* Prestin is required for electromotility of the outer hair cell and for the cochlear amplifier. *Nature* **419**, 300–304 (2002).
- Brownell, W. E., Bader, C. R., Bertrand, D. & de Ribaupierre, Y. Evoked mechanical responses of isolated cochlear outer hair cells. *Science* **227**, 194–196 (1985).
- Ashmore, J. F. A fast motile response in guinea-pig outer hair cells: the cellular basis of the cochlear amplifier. *J. Physiol. (Lond.)* **388**, 323–347 (1987).
- Hudspeth, A. J. Mechanical amplification of stimuli by hair cells. *Curr. Opin. Neurobiol.* **7**, 480–486 (1997).
- Fettiplace, R., Ricci, A. J. & Hackney, C. M. Clues to the cochlear amplifier from the turtle ear. *Trends Neurosci.* **24**, 169–175 (2001).
- Hudspeth, A. J. How the ear's works work. *Nature* **341**, 397–404 (1989).
- Kros, C. J., Rüsch, A. & Richardson, G. P. Mechano-electrical transducer currents in hair cells of the cultured neonatal mouse cochlea. *Proc. R. Soc. Lond. B* **249**, 185–193 (1992).
- Kennedy, H. J., Evans, M. G., Crawford, A. C. & Fettiplace, R. Fast adaptation of mechano-electrical transducer channels in mammalian cochlear hair cells. *Nature Neurosci.* **6**, 832–836 (2003).
- He, D. Z. Z., Jia, S. & Dallos, P. Mechano-electrical transduction of adult outer hair cells studied in a gerbil cochlea. *Nature* **429**, 766–770 (2004).
- Eatock, R. A. Adaptation in hair cells. *Annu. Rev. Neurosci.* **23**, 285–314 (2000).
- Ricci, A. J., Crawford, A. C. & Fettiplace, R. Active hair bundle motion linked to fast transducer adaptation in auditory hair cells. *J. Neurosci.* **20**, 7131–7142 (2000).
- Crawford, A. C. & Fettiplace, R. The mechanical properties of ciliary bundles of turtle cochlear hair cells. *J. Physiol. (Lond.)* **364**, 359–379 (1985).
- Howard, J. & Hudspeth, A. J. Compliance of the hair bundle associated with gating of mechano-electrical transduction channels in the bullfrog's saccular hair cell. *Neuron* **1**, 189–199 (1988).
- Ricci, A. J., Crawford, A. C. & Fettiplace, R. Mechanisms of active hair bundle motion in auditory hair cells. *J. Neurosci.* **22**, 44–52 (2002).
- Russell, I. J., Kössel, M. & Richardson, G. P. Nonlinear mechanical responses of mouse cochlear hair bundles. *Proc. R. Soc. Lond. B* **250**, 217–227 (1992).
- van Netten, S. M. & Kros, C. J. Gating energies and forces of the mammalian hair cell transducer channel and related hair bundle mechanics. *Proc. R. Soc. Lond. B* **267**, 1915–1923 (2000).
- Martin, P., Mehta, A. D. & Hudspeth, A. J. Negative hair-bundle stiffness betrays a mechanism for mechanical amplification by the hair cell. *Proc. Natl Acad. Sci. USA* **97**, 12026–12031 (2000).
- Ricci, A. J., Wu, Y. C. & Fettiplace, R. The endogenous calcium buffer and the time course of transducer adaptation in auditory hair cells. *J. Neurosci.* **18**, 8261–8277 (1998).
- Strelieff, D. & Flock, A. Stiffness of sensory-cell hair bundles in the isolated guinea pig cochlea. *Hear. Res.* **15**, 19–28 (1984).
- Géléoc, G. S., Lennan, G. W., Richardson, G. P. & Kros, C. J. A quantitative comparison of mechano-electrical transduction in vestibular and auditory hair cells of neonatal mice. *Proc. R. Soc. Lond. B* **264**, 611–621 (1997).
- Langer, M. G. *et al.* Lateral mechanical coupling of stereocilia in cochlear hair bundles. *Biophys. J.* **80**, 2608–2621 (2001).
- Rzadzinska, A., Schneider, M. E., Davies, C., Riordan, G. P. & Kachar, B. An actin molecular treadmill and myosins maintain stereocilia functional architecture and self-renewal. *J. Cell Biol.* **164**, 887–897 (2004).
- Kros, C. J. *et al.* Reduced climbing and increased slipping adaptation in cochlear hair cells of mice with Myo7a mutations. *Nature Neurosci.* **5**, 41–47 (2002).
- Müller, M. Frequency representation in the rat cochlea. *Hear. Res.* **51**, 247–254 (1991).
- Assad, J. A. & Corey, D. P. An active motor model for adaptation by vertebrate hair cells. *J. Neurosci.* **12**, 3291–3309 (1992).
- Dallos, P. & Evans, B. N. High frequency outer hair cell motility: corrections and corrigendum. *Science* **268**, 1420–1421 (1995).
- Santos-Sacchi, J. On the frequency limit and phase of outer hair cell motility: effects of the membrane filter. *J. Neurosci.* **12**, 1906–1916 (1992).
- Preyer, P., Renz, S., Hemmert, W., Zenner, H.-P. & Gummer, A. W. Receptor potential of outer hair cells isolated from base to apex of the adult guinea pig cochlea: implications for cochlear tuning mechanisms. *Aud. Neurosci.* **2**, 145–157 (1996).

Acknowledgements This work was supported by a grant to R.F. from the National Institutes on Deafness and other Communicative Disorders (NIH).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.F. (fettiplace@physiology.wisc.edu).

Pax3 functions at a nodal point in melanocyte stem cell differentiation

Deborah Lang^{1*}, Min Min Lu¹, Li Huang¹, Kurt A. Engleka¹, Maozhen Zhang¹, Emily Y. Chu², Shari Lipner³, Arthur Skoultschi³, Sarah E. Millar² & Jonathan A. Epstein¹

¹Cardiovascular Division, Department of Medicine, and ²Department of Dermatology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA
³Department of Cell Biology, Albert Einstein College of Medicine, Bronx, New York 10461, USA

* Present address: Section of Dermatology, Department of Medicine, University of Chicago, 5841 South Maryland Avenue, MC 5067, L504, Chicago, Illinois 60637, USA

Most stem cells are not totipotent. Instead, they are partially committed but remain undifferentiated. Upon appropriate stimulation they are capable of regenerating mature cell types¹. Little is known about the genetic programmes that maintain the undifferentiated phenotype of lineage-restricted stem cells. Here we describe the molecular details of a nodal point in adult melanocyte stem cell differentiation in which Pax3 simultaneously functions to initiate a melanogenic cascade while acting downstream to prevent terminal differentiation. Pax3 activates expression of *Mitf*, a transcription factor critical for melanogenesis^{2,3}, while at the same time it competes with *Mitf* for occupancy of an enhancer required for expression of dopachrome tautomerase, an enzyme that functions in melanin synthesis⁴. Pax3-expressing melanoblasts are thus committed but undifferentiated until Pax3-mediated repression is relieved by activated β -catenin. Thus, a stem cell transcription factor can both determine cell fate and simultaneously maintain an undifferentiated state, leaving a cell poised to differentiate in response to external stimuli.

We identified an unappreciated expression domain for Pax3 in hair follicles (Fig. 1a–d). Pax3 immunoreactive cells in skin also express β -galactosidase in *Dct-lacZ* mice (Fig. 1e) as well as endogenous dopachrome tautomerase^{5,6} (*Dct*) (Fig. 1f), and these cells are located in the bulge region of mature hair follicles where melanocyte stem cells are found⁷ (Fig. 1g). We created mice in which the endogenous Pax3 locus was modified to express Cre recombinase and crossed them to β -galactosidase reporter mice to confirm the existence of Pax3 descendants in the hair follicle (Fig. 1h). Pax3 is expressed by label-retaining cells in resting, telogen stage adult follicles (Fig. 1i–k), further supporting their identity as stem cells⁷. Some Pax3-expressing cells within the hair follicle do not express *Dct-lacZ* or *Dct* (arrows, Fig. 1e) suggesting that these cells may not yet have initiated *Dct* expression. We examined Pax3 expression in skin from mice in which *Wnt1*-expressing neural crest derivatives are labelled with green fluorescent protein (GFP). Pax3 expression colocalizes with GFP suggesting that Pax3-positive cells derive from neural crest (Fig. 1l–n). The majority of Pax3-expressing cells in early anagen stage newborn follicles also express *Mitf* (ref. 8; Fig. 1o–q) and *Sox10* (ref. 9; Fig. 1r–t).

We tested the ability of Pax3, *Sox10* or *Mitf*, alone or in combination, to activate expression of a reporter construct containing a 3,181-base-pair (bp) *Dct* genomic fragment previously used to create *Dct-lacZ* mice. This fragment includes four regions of conservation (>75%) between mouse and human genomes (Fig. 2a). *Sox10* and *Mitf* induce synergistic activation of reporter activity of over 100-fold. The addition of Pax3 to *Sox10* and/or *Mitf* results in significant repression (Fig. 2a). The region between –80 and –350 bp upstream of the *Dct* transcription start (enhancer region 2, Fig. 2a) contains a functional Pax3, *Sox10* and *Mitf* responsive element.

Pax3, *Sox10* and *Mitf* are each able to associate with the enhancer as determined by chromatin immunoprecipitation (ChIP) (Supplementary Fig. S1a). Pax3 directly activates expression of *Mitf* by

functioning synergistically with *Sox10* (refs 2, 10). We compared the ability of Pax3 to regulate expression of *Mitf* (ref. 3) and *Dct* reporter constructs. Under conditions in which Pax3 activates expression of *Mitf*, identical concentrations of Pax3 repress the *Dct* reporter (Fig. 2b).

Potential *Mitf* and *Sox10* binding sites are present within enhancer region 2 (Fig. 2f). The putative *Mitf* binding sequence, or M-box^{11,12}, has been reported previously and an adjacent atypical *Lef*/*Tcf* binding site contributes to *Mitf*-mediated *Dct* regulation⁴. Pax3 is able to bind with relatively low affinity to the enhancer sequence (Supplementary Fig. S1b). Mutation of the M-box destroys the ability of Pax3 to bind to the enhancer (Supplementary Fig. S1b)

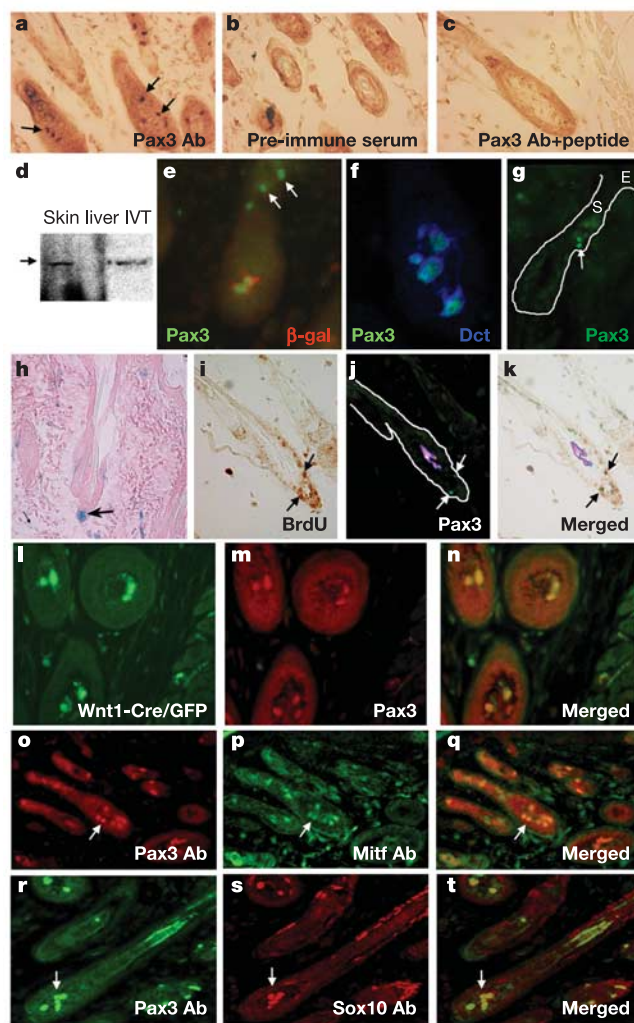


Figure 1 Pax3 is expressed in mature hair follicles. **a–c**, Pax3 expression in early anagen hair follicles of 2-day-old mice is visualized using Pax3 antibody (Ab) (**a**, arrows) in comparison with controls using pre-immune serum (**b**) or antibody pre-adsorbed with Pax3 peptide (**c**). **d**, Immunoprecipitation and western analysis for Pax3 reveals a 68-kD protein in skin but not liver. *In vitro* translated (IVT) Pax3 is used as positive control (lane 3). **e–f**, Pax3 (green) is co-expressed with β -galactosidase driven from a *Dct* promoter (**e**, red) or endogenous *Dct* (**f**, blue) detected by immunohistochemistry in newborn skin. Pax3 is also expressed in the follicle independently of *Dct* (**e**, arrows). **g**, Pax3-expressing cells (green, arrow) in resting telogen stage adult hair follicles. S, shaft. E, epithelial surface. **h**, Pax3-Cre knockin mice were used to activate expression of β -galactosidase (blue, arrow). **i–k**, Pax3 is co-expressed with label-retaining cells, identified by 5-bromodeoxyuridine (BrdU) incorporation and long-term retention (**i**) in resting telogen stage follicles of 58-day-old mice. Pax3 is green (**j**) and an overlay of **i** and **j** is shown in **k**. **l–n**, *Wnt1*-Cre transgenic mice crossed into a Z/EG background label neural crest derivatives by expression of GFP (**l**, green), which colocalizes with Pax3 (**m**). (Overlay shown in **n**.) Pax3 expression overlaps with *Mitf* (**o–q**, arrows) and *Sox10* (**r–t**, arrows) in early anagen P2 follicles.

and to repress basal reporter gene activity (Fig. 2c). The ability of Pax3 to repress basal activity or Mitf-dependent activation is at least partially dependent upon an intact Pax3 paired type DNA binding domain because a mutation that abolishes paired domain DNA binding (Pro to Leu at amino acid 50) diminishes repressor function (Supplementary Fig. S1d). A functional Sox10 binding site is downstream of the M-box (Supplementary Fig. S1c).

These observations suggest that Pax3 and Mitf compete for binding to the *Dct* enhancer. Saturation curves for Mitf activation and ChIP assays are consistent with Pax3 acting as a competitive inhibitor (Fig. 2d, e). When transfected at equal ratios, or with a slight excess of Mitf expression plasmid, Pax3 preferentially associates with enhancer DNA. At high ratios of Mitf:Pax3, Mitf displaces Pax3. Hence, Pax3, functioning with Sox10 (refs 2, 10), can activate expression of *Mitf* while simultaneously acting as a competitive inhibitor of Mitf-mediated activation of *Dct*.

Mutation of the Lef/Tcf binding site abolishes Pax3-mediated repression (Fig. 2g). Pax5, a factor closely related to Pax3, is able to physically interact with Lef1 (ref. 13) and both Pax5 and Pax2 can recruit Groucho co-repressors^{14,15}. Lef/Tcf factors can interact with Groucho co-repressors¹⁶. Pax3 is able to recruit Grg4 (Fig. 2h), which is expressed in Dct- and Pax3-positive cells (Fig. 2i, j). Pax3 and Grg4 physically interact (Supplementary Fig. S1e). Endogenous Grg4 is present in 293T cells and could account for Pax3-mediated

repression observed in that cell line (Supplementary Fig. S1f).

Lef/Tcfs are cofactors for nuclear β -catenin and mediate canonical Wnt signalling¹⁷. Co-transfection of activated β -catenin abolishes repressor activity of Pax3 (Fig. 3a) and displaces Pax3 from *Dct* enhancer DNA (Fig. 3b) thus preventing competition with Mitf (Fig. 3c). Activated β -catenin displaces Grg4 from the *Dct* enhancer (Fig. 2h), explaining the loss of Pax3-mediated repression. Lef1, Pax3 and Grg4 can form a complex in solution, and activated β -catenin displaces Pax3 (Fig. 3d).

B16 melanoma cells harbour an activating mutation in β -catenin and express Pax3, Mitf, Sox10, Grg4 and Dct. Pax3 is not located at the endogenous *Dct* enhancer in B16 cells (Fig. 3e). Transfection of dominant negative Lef1 induces endogenous Pax3 and Grg4 proteins to occupy the endogenous *Dct* enhancer (Fig. 3e) and levels of Dct RNA (data not shown) and protein (Fig. 3f) are reduced. Because β -catenin also regulates *Mitf*, we confirmed that the effects of dominant negative Lef1 were maintained in the presence of excess Mitf (Fig. 3e, f).

We examined Pax3 and Dct expression in skin from TOPGAL mice, which serve as a reporter for activated β -catenin signalling¹⁸. In newborn anagen stage hair follicles, cells that express both Pax3 and Dct also express β -galactosidase (Fig. 4a–c). However, when Pax3 is present in the absence of activated β -catenin signalling, Dct is not expressed (Fig. 4d–f). We expressed a soluble inhibitor of Wnt

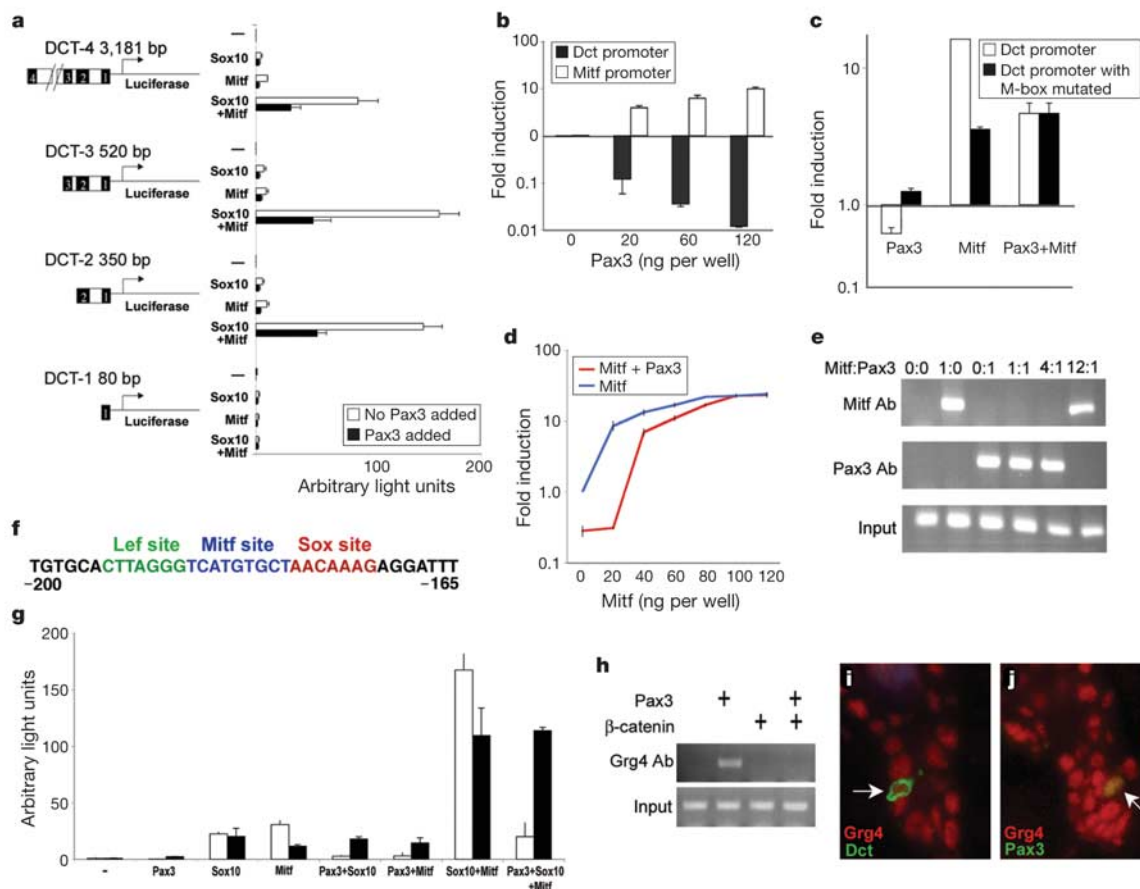


Figure 2 Pax3 and Mitf regulate *Dct* and compete for enhancer occupancy. **a**, Deletion analysis of the *Dct* upstream sequence. Homology between human and mouse is indicated (black boxes, numbered 1–4). Transfections in 293T cells are without (white bars) or with (black bars) Pax3 protein. All transfections are expressed as the average of at least three independent experiments in triplicate, $\pm$ s.d. **b**, Pax3 represses the *Dct* reporter while activating the *Mitf* reporter. **c**, Pax3 and/or Mitf is transfected with DCT-2 reporter (white bars) or with DCT-2 reporter with M-box mutated (black bars). **d**, Saturation curve for Mitf-mediated activation of DCT-2, with (red line) or without (blue line) 20 ng Pax3 expression construct. **e**, ChIP analysis of 293T cells transfected with Mitf

and/or Pax3. Amplified DNA is from homology region 2. Primers for β -actin and luciferase failed to yield product. **f**, Sequence of *Dct* enhancer. **g**, Pax3, Mitf and/or Sox10 is transfected with DCT-2 reporter construct (white bars) or with DCT-2 reporter with the Lef1 site mutated (black bars). **h**, ChIP analysis of 293T cells transfected with Pax3 and/or activated β -catenin using Grg4 antibody to immunoprecipitate endogenous Grg4. Primers for β -actin and luciferase failed to yield product. **i**, Immunohistochemistry of 2-month-old mouse resting hair follicle using Grg4 (red) and Dct (green) antibodies. **j**, Immunohistochemistry of 2-month-old follicle using Pax3 (green) and Grg4 (red) antibodies.

signalling using an inducible system^{19,20}. Consistent with previous data^{21–23}, we observed a marked decrease in Pax3-expressing cells in the skin of transgenic mice induced to express Dickkopf 1 (Dkk1). However, some Pax3-expressing cells were observed after late embryonic induction of Dkk1. In wild-type embryonic skin, the majority of Pax3-expressing cells also express Dct (98% of 200 cells analysed). In Dkk1-overexpressing embryos, the percentage of Pax3-expressing cells that also express Dct is markedly decreased (38% of 180 cells analysed).

We generated mouse embryos in which β -catenin was inactivated in Pax3-expressing cells²⁴. We confirmed loss of β -catenin by immunohistochemistry, and homozygous floxed β -catenin embryos that carried the Pax3-Cre allele were identified at mid-gestation with neural tube defects (data not shown). At embryonic day 13.5 (E13.5), loss of β -catenin in Pax3-expressing cells resulted in the loss of Dct expression in skin, in accord with related previous work^{25,26}. However, we were able to identify Pax3-expressing cells in the skin of these embryos, albeit in fewer numbers than in wild-type litter mates where we found many Pax3-expressing cells that also expressed Mitf and Dct (Fig. 4g–i). In skin lacking β -catenin in Pax3 cells, Pax3 and Mitf were co-expressed, but Dct was absent (Fig. 4j–l). Hence, β -catenin is required for Dct expression in Pax3-expressing melanocyte precursors.

Our data suggest a model to explain how expression of an upstream determination factor, such as Pax3, might initiate a lineage-specific gene programme while at the same time prevent terminal differentiation. In melanocyte precursors, we suggest that Pax3 functions with Sox10 to activate *Mitf* expression, while at the same time it prevents Mitf from activating downstream genes (Fig. 4m). This nodal checkpoint is characterized by Pax3 compe-

tion with Mitf for enhancer occupancy (Fig. 4n). Thus, the Pax3-expressing precursor is unable to fully differentiate, while Mitf is able to accumulate, resulting in a 'biological capacitor' in which the cell is primed to rapidly express downstream genes once Pax3-mediated repression is relieved. We postulate that external stimuli that result in melanocyte stem cell activation, such as injury or sun exposure, function through activation of β -catenin, resulting in displacement of Pax3 and associated Groucho co-repressors and activation of downstream gene expression.

The ability of a single factor, or complex of factors, to simultaneously activate a determination programme while preventing terminal differentiation may represent a general paradigm for developmental and stem cell biology. This paradigm predicts a class of 'pangenes' that encode related functions, reminiscent of the Greek god Pan, and Peter Pan, who were able to orchestrate complex events while never growing to maturity. For instance, Sox10 maintains neural crest lineage multipotency while inhibiting terminal neuronal differentiation²⁷. During eye development, the conserved *Pax6* gene is required for specification of multiple ocular lineages. Loss of Pax6 expression in *Drosophila*, mice and man results in a complete absence of eye formation because critical developmental

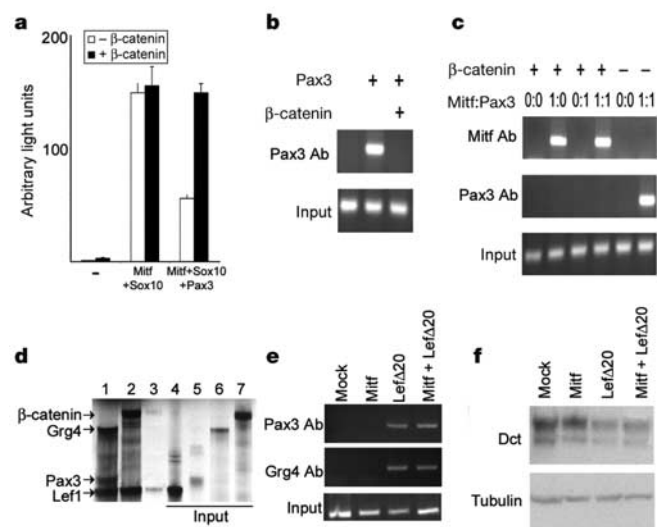


Figure 3 Activated β -catenin modulates Pax3 activity. **a**, DCT-2 reporter is transfected alone, with Mitf and Sox10 or with Mitf, Sox10 and Pax3 without (white bars) or with (black bars) activated β -catenin. **b**, **c**, ChIP analysis from 293T cells transfected with Mitf, Pax3 and/or activated β -catenin. Primers for β -actin and luciferase failed to yield product. β -catenin displaces Pax3 from DCT-2 enhancer DNA (**b**) and β -catenin prevents Pax3 from effectively competing with Mitf for occupancy of the *Dct* enhancer site (**c**). **d**, *In vitro* translated Grg4, Pax3 and Lef1 proteins were incubated together and immunoprecipitated with anti-Lef1 antibody, which resulted in precipitation of all three proteins (lane 1). Further addition of β -catenin resulted in loss of Pax3 and Grg4 in the immunoprecipitate (lane 2). Lane 3 is negative control immunoprecipitate in the absence of Lef1 antibody. Lanes 4–7 are input Lef1, Pax3, Grg4 and β -catenin proteins, respectively. **e**, ChIP analysis in B16 melanoma cells with primers for endogenous *Dct* enhancer. Endogenous Pax3 and Grg4 proteins occupy the *Dct* enhancer when a mutant Lef1 protein that does not bind to β -catenin (Lef Δ 20) is present. β -actin genomic primers, negative control, provided no amplification of DNA (not shown). **f**, Western analysis of B16 cells transfected without or with dominant negative Lef1 (Lef Δ 20).

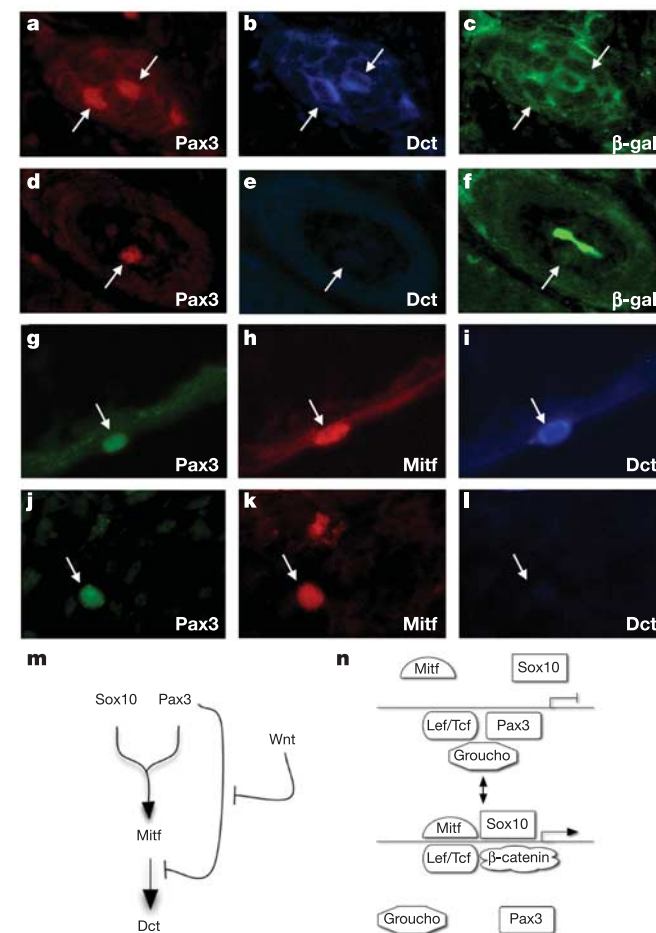


Figure 4 Dct expression requires activated β -catenin. **a–f**, Pax3, Dct and β -galactosidase expression in anagen stage hair follicles of 2-day-old TOPGAL mice. Cells in the follicles express all three proteins (**a–c**) or only Pax3 (**d–f**). **g–i**, Pax3, Mitf and Dct expression in skin of E13.5 wild-type embryos (**g–i**) or embryos lacking β -catenin in Pax3-expressing cells (**j–l**). Occasional Pax3-expressing cells, which co-express Mitf in skin, are identified in tissue-specific β -catenin mutants, but Dct expression is undetectable in these cells (**i–l**). **m, n**, Model depicting the ability of Pax3 to activate a melanogenic cascade while simultaneously competing with Mitf for activation of *Dct*, thus preventing expression of terminal differentiation markers until external stimuli abolish Pax-mediated repression.

programmes are never initiated²⁸. However, inactivation of Pax6 later in development, at the retinal progenitor stage, results in loss of ability of this committed but undifferentiated cell type to maintain pluripotentiality²⁹. Pax5, which can also interact with Grg4 (ref. 14), may play similar roles in lymphocyte development³⁰.

Adult resident stem cells have been identified in a large number of organs and provide exciting potential for tissue regeneration. A fundamental understanding of the molecular programmes regulating both differentiation and maintenance of the undifferentiated state will be required to harness this potential. Our work characterizes a critical regulatory circuit that exemplifies conservation of genetic programmes between embryonic neural crest development and adult melanocyte stem cell function. Additional nodal checkpoints, with parallel transcriptional circuits, are likely to exist in other embryonic and adult stem cells. □

Methods

Immunohistochemistry

Immunohistochemistry was performed on paraffin-embedded tissue fixed in 4% paraformaldehyde. Antigen was exposed using Bull's Eye reagent (Biocare Medical) and heated in a pressure cooker. Antibodies utilized were Pax3 (polyclonal sera or monoclonal supernatant, Developmental Studies Hybridoma Bank, 1:3,000 for DAB staining, 1:800 for immunofluorescence), Mitf (Vector Laboratories, 1:10), Sox10 (Chemicon International, 1:20), β -galactosidase (Promega Corporation, 1:100), Dct/Trp2 and Grg4 (Santa Cruz Biotechnology, 1:50 and 1:100, respectively). Secondary antibodies conjugated with fluorescent tags (Alexa Fluor, Molecular Probes) were used at a dilution of 1:250. For label retention studies, mice were injected subcutaneously with BrdU (10 μ g per g body weight) twice daily from P20 to P27. Skin was collected at P58. BrdU antibody (Biocare Medical) was used at a dilution of 1:200.

Cell culture, transfection and ChIP assays

293T cells and B16 cells (American Type Culture Collection) were maintained in DMEM supplemented with 10% fetal bovine serum (Invitrogen Life Technologies). A total of 0.5 μ g of DNA was mixed with 10 μ l Effectene (Qiagen). Luciferase activity (Luciferase assay kit, Promega Corporation) was normalized for transfection efficiency using pCMV β (BD Biosciences/Clontech) and expressed as either fold activation compared with reporter construct alone, or as arbitrary light units. For ChIP assays, transfected cells were fixed in 1% formaldehyde and quenched in 0.125 M glycine, then processed according to the manufacturer's protocol (Upstate Biotechnology). Polymerase chain reaction (PCR) was performed with primers to the Dct enhancer region-2 GGAGAAGTACTTAGCAATGCAC AGG (F) and AGCCATCATTAAGGGGATTATAACC (R). All ChIP samples were tested for false positive PCR amplification using primers that amplify sequence from the β -actin gene (for genomic DNA contamination) and luciferase (reporter construct contamination). In all cases, these amplifications failed to yield product.

Details of methods for electrophoretic mobility shift assays, immunoprecipitation, western blotting, and constructs and mouse lines used are provided in the Supplementary Methods.

Received 19 October; accepted 10 December 2004; doi:10.1038/nature03292.

- Fuchs, E., Tumber, T. & Guasch, G. Socializing with the neighbors: stem cells and their niche. *Cell* **116**, 769–778 (2004).
- Potterf, S. B., Furumura, M., Dunn, K. J., Arnheiter, H. & Pavan, W. J. Transcription factor hierarchy in Waardenburg syndrome: regulation of MITF expression by SOX10 and PAX3. *Hum. Genet.* **107**, 1–6 (2000).
- Watanabe, A., Takeda, K., Ploplis, B. & Tachibana, M. Epistatic relationship between Waardenburg syndrome genes MITF and PAX3. *Nature Genet.* **18**, 283–286 (1998).
- Yasumoto, K. *et al.* Microphthalmia-associated transcription factor interacts with LEF-1, a mediator of Wnt signaling. *EMBO J.* **21**, 2703–2714 (2002).
- Steel, K. P., Davidson, D. R. & Jackson, I. J. TRP-2/DT, a new early melanoblast marker, shows that steel growth factor (c-kit ligand) is a survival factor. *Development* **115**, 1111–1119 (1992).
- Tsukamoto, K., Jackson, I. J., Urabe, K., Montague, P. M. & Hearing, V. J. A second tyrosinase-related protein, TRP-2, is a melanogenic enzyme termed DOPachrome tautomerase. *EMBO J.* **11**, 519–526 (1992).
- Nishimura, E. K. *et al.* Dominant role of the niche in melanocyte stem-cell fate determination. *Nature* **416**, 854–860 (2002).
- Widlund, H. R. & Fisher, D. E. Microphthalmia-associated transcription factor: a critical regulator of pigment cell development and survival. *Oncogene* **22**, 3035–3041 (2003).
- Kuhlbrodt, K., Herbarth, B., Sock, E., Hermans-Borgmeyer, I. & Wegner, M. Sox10, a novel transcriptional modulator in glial cells. *J. Neurosci.* **18**, 237–250 (1998).
- Bondurand, N. *et al.* Interaction among SOX10, PAX3 and MITF, three genes altered in Waardenburg syndrome. *Hum. Mol. Genet.* **9**, 1907–1917 (2000).
- Bertolotto, C. *et al.* Microphthalmia gene product as a signal transducer in cAMP-induced differentiation of melanocytes. *J. Cell Biol.* **142**, 827–835 (1998).
- Aksan, I. & Goding, C. R. Targeting the microphthalmia basic helix-loop-helix-leucine zipper transcription factor to a subset of E-box elements *in vitro* and *in vivo*. *Mol. Cell. Biol.* **18**, 6930–6938 (1998).
- Jin, Z. X. *et al.* Lymphoid enhancer-binding factor-1 binds and activates the recombination-activating gene-2 promoter together with c-Myb and Pax-5 in immature B cells. *J. Immunol.* **169**, 3783–3792 (2002).

- Eberhard, D., Jimenez, G., Heavey, B. & Busslinger, M. Transcriptional repression by Pax5 (BSAP) through interaction with corepressors of the Groucho family. *EMBO J.* **19**, 2292–2303 (2000).
- Cai, Y., Brophy, P. D., Levitan, I., Stifani, S. & Dressler, G. R. Groucho suppresses Pax2 transactivation by inhibition of JNK-mediated phosphorylation. *EMBO J.* **22**, 5522–5529 (2003).
- Roose, J. *et al.* The *Xenopus* Wnt effector XTcf-3 interacts with Groucho-related transcriptional repressors. *Nature* **395**, 608–612 (1998).
- Behrens, J. *et al.* Functional interaction of beta-catenin with the transcription factor LEF-1. *Nature* **382**, 638–642 (1996).
- DasGupta, R. & Fuchs, E. Multiple roles for activated LEF/TCF transcription complexes during hair follicle development and differentiation. *Development* **126**, 4557–4568 (1999).
- Chu, E. Y. *et al.* Canonical WNT signaling promotes mammary placode development and is essential for initiation of mammary gland morphogenesis. *Development* **131**, 4819–4829 (2004).
- Andl, T., Reddy, S. T., Gaddapara, T. & Millar, S. E. WNT signals are required for the initiation of hair follicle development. *Dev. Cell* **2**, 643–653 (2002).
- Ikeya, M., Lee, S. M., Johnson, J. E., McMahon, A. P. & Takada, S. Wnt signalling required for expansion of neural crest and CNS progenitors. *Nature* **389**, 966–970 (1997).
- Lee, H. Y. *et al.* Instructive role of Wnt/beta-catenin in sensory fate specification in neural crest stem cells. *Science* **303**, 1020–1023 (2004).
- Garcia-Castro, M. I., Marcelle, C. & Bronner-Fraser, M. Ectodermal Wnt function as a neural crest inducer. *Science* **297**, 848–851 (2002).
- Huelsken, J., Vogel, R., Erdmann, B., Cotsarelis, G. & Birchmeier, W. Beta-catenin controls hair follicle morphogenesis and stem cell differentiation in the skin. *Cell* **105**, 533–545 (2001).
- Hari, L. *et al.* Lineage-specific requirements of beta-catenin in neural crest development. *J. Cell Biol.* **159**, 867–880 (2002).
- Braut, V. *et al.* Inactivation of the beta-catenin gene by Wnt1-Cre-mediated deletion results in dramatic brain malformation and failure of craniofacial development. *Development* **128**, 1253–1264 (2001).
- Kim, J., Lo, L., Dormand, E. & Anderson, D. J. SOX10 maintains multipotency and inhibits neuronal differentiation of neural crest stem cells. *Neuron* **38**, 17–31 (2003).
- Ashery-Padan, R. & Gruss, P. Pax6 lights-up the way for eye development. *Curr. Opin. Cell Biol.* **13**, 706–714 (2001).
- Marquardt, T. *et al.* Pax6 is required for the multipotent state of retinal progenitor cells. *Cell* **105**, 43–55 (2001).
- Nutt, S. L., Heavey, B., Rolink, A. G. & Busslinger, M. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* **401**, 556–562 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Glick for K5-rtTA mice, M. Shin and E. Morrissey for mice and scientific advice, and T. Andl, A. Souabni, C. Lobe, W. Birchmeier, G. Oliver, T. Force and P. Hamel for reagents. This work was supported by grants from the NIH to S.E.M. and J.A.E.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.A.E. (epsteinj@mail.med.upenn.edu).

Toll-like receptor 3 promotes cross-priming to virus-infected cells

Oliver Schulz^{1*}, Sandra S. Diebold^{1*}, Margaret Chen^{2,3}, Tanja I. Näslund², Martijn A. Nolte¹, Lena Alexopoulou^{4,†}, Yasu-Taka Azuma⁴, Richard A. Flavell⁴, Peter Liljestrom^{2,3} & Caetano Reis e Sousa¹

¹Immunobiology Laboratory, Cancer Research UK London Research Institute, Lincoln's Inn Fields Laboratories, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

²Microbiology and Tumor Biology Center, Karolinska Institutet, Nobelsväg 16, Solna, SE-171 77 Stockholm, Sweden

³Department of Vaccine Research at Swedish Institute for Infectious Disease Control, Nobelsväg 18, 171 77 Stockholm, Sweden

⁴Section of Immunobiology, Yale University School of Medicine and Howard Hughes Medical Institute, 330 Cedar Street, PO Box 208011, New Haven, Connecticut 06520, USA

* These authors contributed equally to this work

† Present address: Centre d'Immunologie de Marseille-Luminy, CNRS-INSERM, Parc Scientifique et Technologique de Luminy – Case 906, 13009 Marseille, France

Cross-presentation of cell-associated antigens plays an important role in regulating CD8⁺ T cell responses to proteins that are not expressed by antigen-presenting cells (APCs)¹. Dendritic cells are the principal cross-presenting APCs *in vivo* and much progress has been made in elucidating the pathways that allow dendritic

cells to capture and process cellular material¹. However, little is known about the signals that determine whether such presentation ultimately results in a cytotoxic T cell (CTL) response (cross-priming) or in CD8⁺ T cell inactivation (cross-tolerance). Here we describe a mechanism that promotes cross-priming during viral infections. We show that murine CD8⁺ dendritic cells are activated by double-stranded (ds)RNA present in virally infected cells but absent from uninfected cells. Dendritic cell activation requires phagocytosis of infected material, followed by signalling through the dsRNA receptor, toll-like receptor 3 (TLR3). Immunization with virus-infected cells or cells containing synthetic dsRNA leads to a striking increase in CTL cross-priming against cell-associated antigens, which is largely dependent on TLR3 expression by antigen-presenting cells. Thus, TLR3 may have evolved to permit cross-priming of CTLs against viruses that do not directly infect dendritic cells.

In mice, plasmacytoid dendritic cells (DCs) and CD8⁺ DCs appear to be the major APC subtypes in priming antiviral CTL²⁻⁶. The role of CD8⁺ DCs in this regard might be a consequence of their unique ability to re-present cell-associated antigens on major histocompatibility complex (MHC) class I and class II^{7,8}; this in turn might be related to their ability to phagocytose dying cells^{9,10}. However, this process *per se* is insufficient to induce CTL priming because cross-presentation of cellular material by CD8⁺ DCs has also been implicated in CD8⁺ T cell tolerance^{11,12}. Therefore, virally infected cells, but not their uninfected counterparts, must generate signals that act on DCs to favour cross-priming. One such signal might be secreted type I interferons (IFN- α/β)¹³. Here, we asked whether other signals are preserved in virally infected cells. Viral infection is accompanied by production of dsRNA, a potent innate stimulus that is recognized, in part, by TLR3 (ref. 14). We have previously found that CD8⁺ DCs express high amounts of

messenger RNA for TLR3 but not for TLR7, a receptor for single stranded RNA prominently expressed by plasmacytoid DCs¹⁵ (see Supplementary Fig. 1a, b). Consistent with those data, exogenous synthetic dsRNA (poly I:C) activates CD8⁺ DCs but not plasmacytoid DCs; the opposite is seen with a TLR7 ligand¹⁵ (see Supplementary Fig. 1c). On the basis of these observations, we hypothesized that TLR3 allows CD8⁺ DCs to 'sense' the presence of dsRNA within virally infected cells and induce cross-priming.

We first determined whether dsRNA activates CD8⁺ DCs when supplied in cell-associated form. Purified CD8⁺ spleen DCs were co-cultured with Vero cells that had been loaded with poly I:C by electroporation (Vero-poly I:C cells) and exposed to ultraviolet light to induce cell death. Both Vero-poly I:C and mock-treated cells were phagocytosed by CD8⁺ DCs, as determined by flow cytometry and confocal microscopy (see Supplementary Fig. 2). However, compared with control cells, Vero-poly I:C cells induced *de novo* expression of *Ifna* and *Ifnb* genes in CD8⁺ DCs and promoted increased expression of mRNA for interleukin-6 (IL-6), tumour-necrosis factor- α (TNF- α), CD40 and CD86 (Fig. 1a). Higher levels of cell-surface expression of CD40, CD86 and CD80 were observed (Fig. 1b), as was the accumulation of IL-6 protein in cell culture supernatants (Fig. 1c). In contrast, IFN- α , IFN- β and TNF- α protein levels remained below the detection limit of the enzyme-linked immunosorbent assay (ELISA), and IL-12 p40 was induced by Vero cells that had not been loaded with poly I:C (Figs 1a and 2a). The ability of poly I:C-loaded cells to activate CD8⁺ DCs was not restricted to Vero cells and could be observed using a variety of cell lines (Fig. 1c), as well as with syngeneic or allogeneic mouse splenocytes (see Supplementary Fig. 3). Ultraviolet irradiation of the target cells was not essential for CD8⁺ DC activation (see Supplementary Fig. 3), perhaps because dsRNA is a pro-apoptotic stimulus¹⁶ or because cell death is not essential for uptake¹⁷.

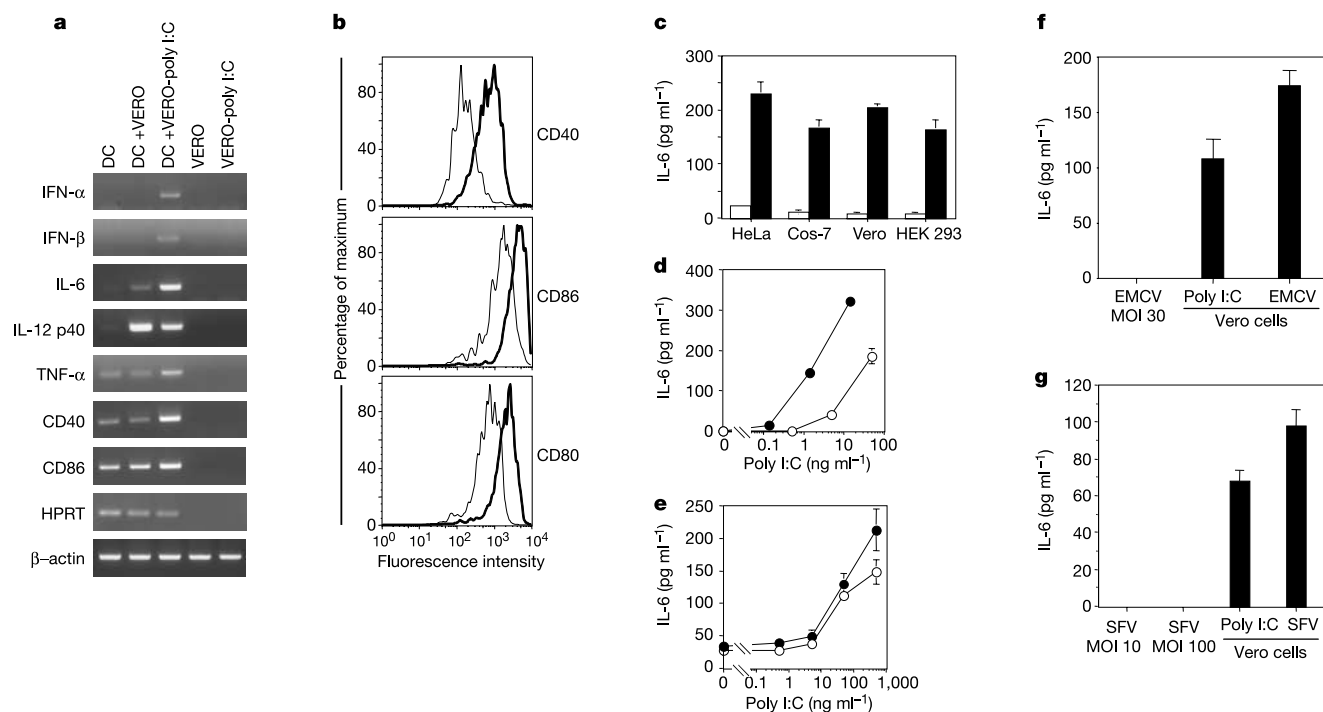


Figure 1 *In vitro* activation of CD8⁺ DCs by poly I:C-loaded or virally infected cells. CD8⁺ DCs, Vero-poly I:C cells or mock-treated Vero cells were cultured in the indicated combinations. **a**, mRNA was analysed by RT-PCR after normalizing for HPRT. All primers except those for β -actin are mouse-specific. **b**, Flow cytometric analysis of co-stimulatory molecule expression by CD8⁺ DCs. Vero-poly I:C cells (thick line), mock-treated Vero cells (thin line). **c–g**, IL-6 accumulation in the cell culture supernatants. **c**, CD8⁺ DCs were cultured with the indicated ultraviolet light-treated cell lines and electroporated in

the presence (black bars) or absence (white bars) of poly I:C. **d**, CD8⁺ DCs were cultured with soluble (open circles) or Vero-associated (filled circles) poly I:C or (**e**) with soluble poly I:C in the presence (filled circles) or absence (open circles) of ultraviolet light-treated Vero cells. **f**, CD8⁺ DCs were cultured with EMCV or EMCV-infected Vero cells. Cells loaded with poly I:C were used as a positive control. **g**, As for (**f**) but using SFV. MOI, multiplicity of infection.

Notably, 20–50-fold less poly I:C was required in cell-associated form than in soluble form to induce CD8 α^+ DC activation (Fig. 1d). This was not due to a synergistic effect of the dsRNA and the cells (Fig. 1e), but might instead reflect differences in the uptake of particulate versus soluble dsRNA, as well as increased local concentration of the stimulus owing to settling of the poly I:C-bearing cells.

To determine whether virally infected cells could similarly activate CD8 α^+ DCs, we first screened for viruses that do not infect CD8 α^+ DCs or induce their activation directly. We found that encephalomyocarditis virus¹⁸ (EMCV) and Semliki Forest virus¹⁹ (SFV) have only a limited ability to infect CD8 α^+ DCs *in vitro* (see Supplementary Fig. 4) and do not induce CD8 α^+ DC activation when added as free viral particles (Fig. 1f, g). In contrast, both viruses infect and replicate efficiently in Vero cells (see Supplementary Fig. 4). Notably, when infected Vero cells were co-cultured with CD8 α^+ DCs, they promoted DC activation as efficiently as poly I:C-loaded controls (Fig. 1f, g). This was not simply a quantitative effect because the infected cells contained less virus than the amount added in free form (data not shown). Furthermore, the SFV used for these experiments is a ‘suicide’ virus, engineered to undergo only one round of infection and producing

no progeny virions (owing to the absence of viral structural genes)²⁰. Thus, viruses that fail to activate DCs when present as free viral particles can stimulate DCs when presented as virally infected cells.

Blocking phagocytosis using actin-stabilizing agents such as latrunculin B, or neutralizing endosomal pH with chloroquine completely prevented CD8 α^+ DC activation by poly I:C-loaded Vero cells (Fig. 2a). The same drugs did not inhibit the poly I:C-independent induction of IL-12 p40 (Fig. 2a), serving as a control for the functional integrity of the cells. These results indicate that CD8 α^+ DC activation is not due to factors released by dsRNA-bearing cells, and suggest instead that phagocytosis of cellular material and subsequent phagosomal acidification are necessary. To identify the pathways involved, we analysed the activation of CD8 α^+ DCs lacking candidate genes involved in viral recognition. The response was independent of protein kinase R (Fig. 2b, left), suggesting that dsRNA was not passively leaking out of DC phagosomes into the cytosol²¹. MyD88, an adaptor downstream of many TLRs but not required for TLR3 signalling, was also not essential (Fig. 2b, right). In contrast, TLR3-deficient CD8 α^+ DCs were completely unable to respond to poly I:C-loaded cells although they responded normally to the control stimulus CpG, which signals via TLR9 (Fig. 2c). Similarly, *Tlr3*^{-/-} CD8 α^+ DCs were no longer activated by EMCV-infected or SFV-infected cells (Fig. 2c). We conclude that TLR3-dependent recognition of dsRNA in phagosomes can mediate activation of CD8 α^+ DCs by poly I:C-loaded or virus-infected cells.

To assess the efficiency with which virally infected cells induce CTL cross-priming, we immunized mice with cells infected with ‘suicide’ SFV encoding a non-secreted form of ovalbumin (OVA) as a model antigen. Because the infected cells cannot release viral progeny (see above), there is no risk of infecting APCs *in vivo* and priming CTLs directly²². As a further precaution, we treated the infected cells with trypsin and succinate, which inactivate any infectious virus particles that remain bound to the cell surface (unpublished observations). As an alternative approach, we also used a virus-free system, immunizing mice with cells loaded with OVA protein in the presence or absence of poly I:C to mimic viral infection. Electroporation resulted in the loading of 50–250 ng OVA per 10⁶ cells, with over 80% of the cells expressing immunodetectable levels of protein (see Supplementary Fig. 5). Co-loading of poly I:C had no effect on antigen levels (see Supplementary Fig. 5a, b). SFV-OVA infection resulted in the expression of 100–200 ng OVA per 10⁶ cells (see Supplementary Fig. 5c), with 30–50% of cells showing positive staining for the protein (Supplementary Fig. 4b). Injecting 10⁶ OVA-electroporated or SFV-OVA-infected cells is therefore equivalent to administering less than 300 ng OVA protein.

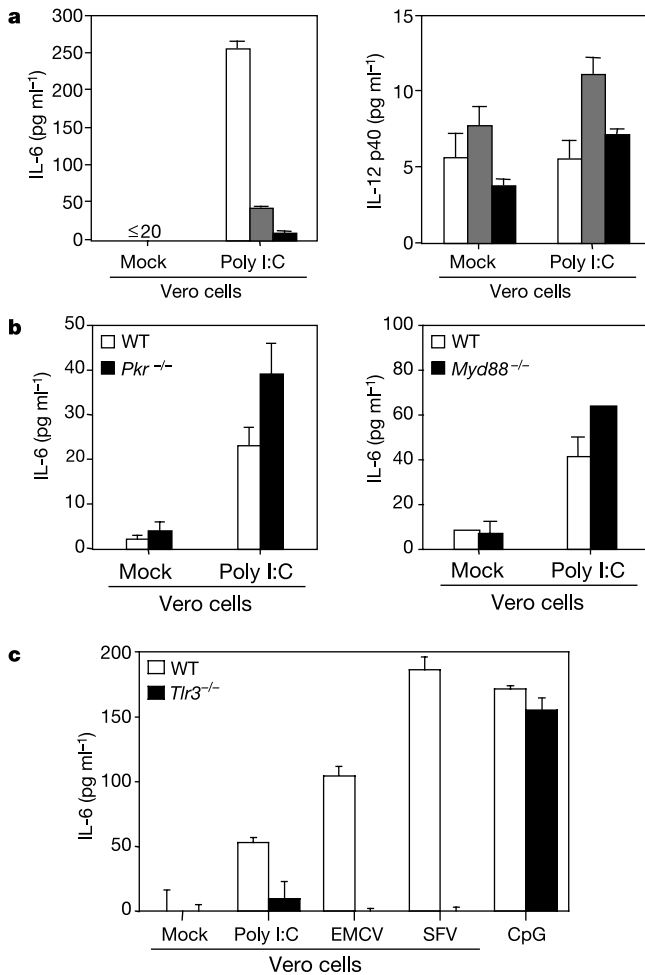


Figure 2 CD8 α^+ DC activation by poly I:C-loaded or virally infected cells requires endosomal recognition by TLR3. **a–c**, Levels of IL-6 and IL-12 p40 in culture supernatants. **a**, CD8 α^+ DCs were co-cultured with poly I:C- or mock-treated Vero cells in the presence of 1 μ M latrunculin B (black), 25 μ M chloroquine (grey), or no inhibitor (white). **b**, CD8 α^+ DCs from *Pkr*^{-/-}, *Myd88*^{-/-} or wild-type (WT) control mice were cultured with poly I:C- or mock-treated Vero cells. **c**, As for (**b**) but comparing WT to *Tlr3*^{-/-} CD8 α^+ DCs and also testing Vero cells infected with EMCV or SFV-OVA. CpG (0.5 μ g ml⁻¹) was used as a control stimulus.

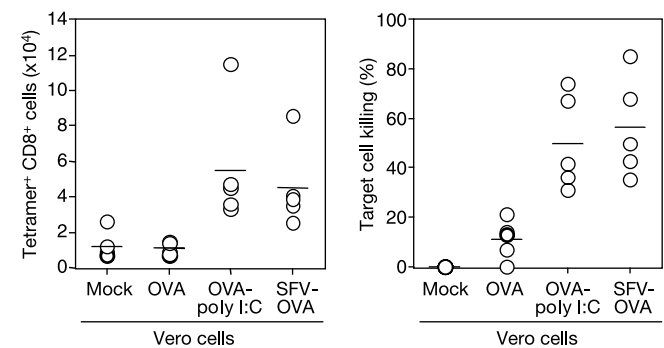


Figure 3 dsRNA or viral infection promotes *in vivo* CTL cross-priming to cell-associated antigen. Naive C57BL/6 mice were immunized with Vero cells loaded with OVA $\pm$ poly I:C or infected with SFV-OVA. Mice were injected with CFSE-labelled target cells on day 6 post-immunization. Splenocytes were isolated the following day and analysed for Thy1.2⁺ tetramer⁺ CD8⁺ cells (left panel) and the persistence of target cells (right panel). Data points represent individual mice.

Remarkably, in cell-associated form with poly I:C, this low level of antigen was sufficient to induce robust cross-priming responses in naïve mice as determined by the expansion of OVA-specific endogenous CD8⁺ T cells or *in vivo* cytotoxicity (Fig. 3 and Supplementary Fig. 6a). Similarly, strong cross-priming of endogenous OVA-specific CTLs was seen after immunization with SFV-OVA-infected Vero cells (Fig. 3 and Supplementary Fig. 6a). In contrast, immunization with Vero cells loaded with OVA alone led to no detectable increase in CTL number and only low levels of OVA-specific *in vivo* cytotoxicity compared with control mice (Fig. 3 and Supplementary Fig. 6a). The adjuvanticity of poly I:C co-loading was also apparent in mice that had received a small number of naïve T cells from OT-I mice to increase the frequency of precursor CTLs before immunization (see Supplementary Fig. 6b). Indeed, the adjuvant effect of cell-associated poly I:C compared favourably with that of anti-CD40 monoclonal antibodies (see Supplementary Fig. 6b), widely regarded as the 'gold standard' for inducing CTL responses. We conclude that virally infected cells are potent inducers of cross-priming *in vivo* and that their immunogenicity can be mimicked by introducing dsRNA into uninfected cells.

To relate the CTL cross-priming data to the TLR3-dependence of CD8α⁺ DC activation, we tested for cross-priming in TLR3-deficient mice. Although these mice mount normal CTL responses

to many viruses²³, they showed impaired responses to SFV-OVA-infected Vero cells as determined by CTL activity or IFN-γ ELISPOT assays (Fig. 4a, b). The impairment in cross-priming was absolute when OVA + poly I:C-loaded cells were used as the immunogen (Fig. 4b). To specifically map the defect in cross-priming to lack of TLR3 expression by APCs, we reconstituted lethally irradiated animals with bone marrow from *Tlr3*^{-/-} donors or *Tlr3*^{+/+} littermate controls, transferred OT-I T cells and immunized with SFV-OVA-infected or OVA + poly I:C-loaded cells. The increased cell immunogenicity conferred by poly I:C loading was completely lost in the *Tlr3*^{-/-} chimaeras (Fig. 4c). Similarly, a reduced response was seen in *Tlr3*^{-/-} chimaeras immunized with SFV-OVA cells (Fig. 4d). As controls, *Tlr3*^{-/-} chimaeras mounted normal responses against OVA-loaded Vero cells co-injected with CpG + anti-CD40 (see Supplementary Fig. 7). Given that both the OT-I cells and the non-haematopoietic compartment in these mice are TLR3-sufficient, these data suggest that TLR3 expression by radiosensitive APCs is involved in *in vivo* cross-priming of CD8⁺ T cells against dsRNA-bearing, virally infected cells.

It is widely accepted that T cell priming requires DC activation by signals associated with the presence of infection. For viruses, direct infection can lead to DC activation via cytosolic pattern recognition of dsRNA intermediates of viral replication²¹. However, viruses that do not have tropism for DCs, such as lymphocytic choriomeningitis

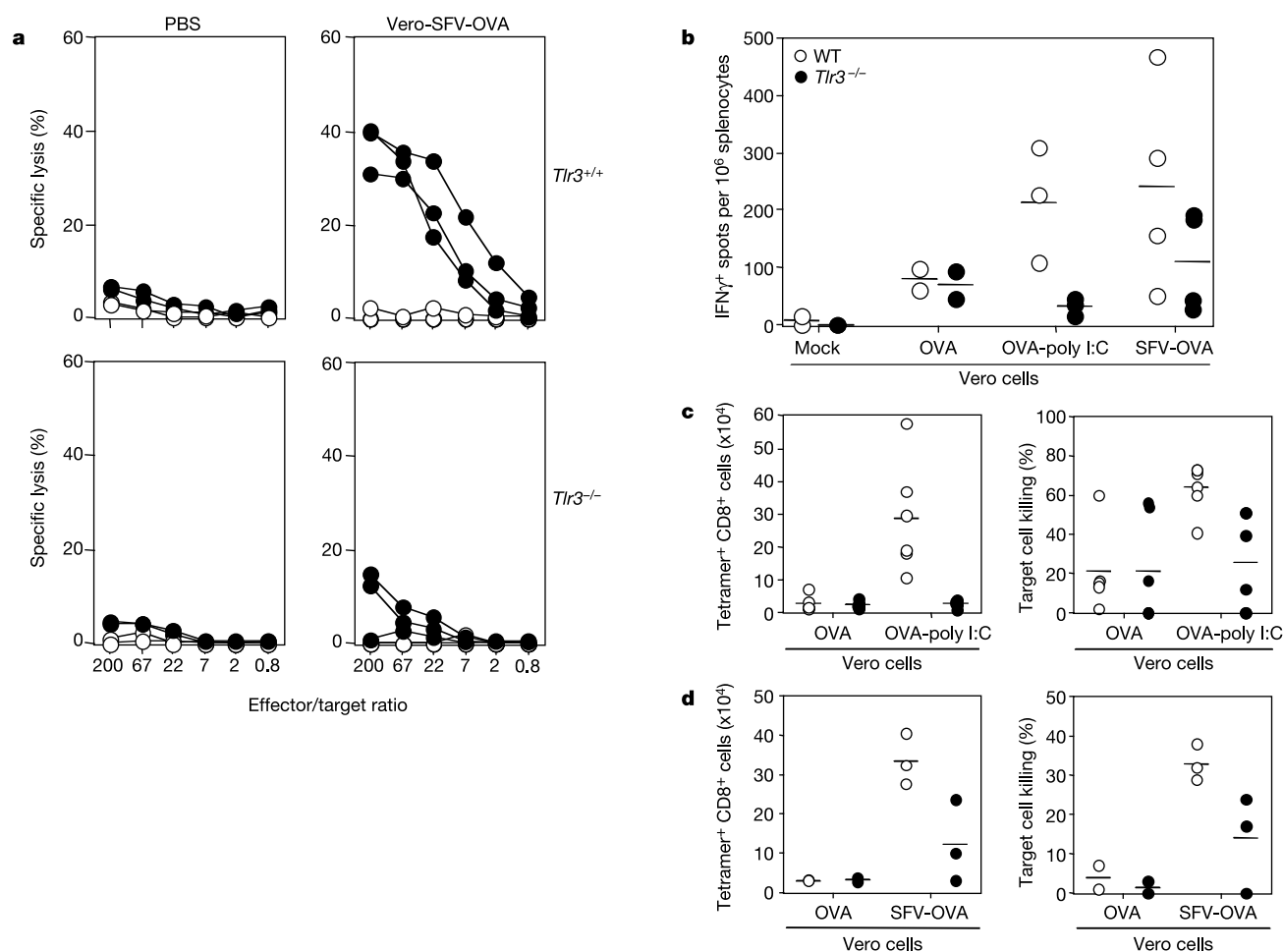


Figure 4 TLR3 dependence of CTL cross-priming against dsRNA-loaded or virally infected cells. **a**, WT (*Tlr3*^{+/+}, upper panels) and *Tlr3*^{-/-} (lower panels) mice were immunized with SFV-OVA-infected Vero cells or PBS. CTL activity was measured by chromium release assay after re-stimulation *ex vivo* using target cells pulsed with OVA peptide (filled circles) or left unpulsed (open circles). **b**, WT (open circles) and *Tlr3*^{-/-} (filled circles) mice were immunized with Vero cells loaded with OVA ± poly I:C, infected with SFV-OVA or mock-

treated. OVA-specific CTLs were measured by IFN-γ ELISPOT on day 9. **c**, **d**, Chimaeric mice reconstituted with bone marrow from *Tlr3*^{+/+} (open circles) or *Tlr3*^{-/-} (filled circles) mice were injected with OT-I T cells and immunized the following day with Vero cells loaded with OVA ± poly I:C (**c**) or Vero cells infected with SFV-OVA (**d**). Numbers of OT-I T cells and *in vivo* specific killing of target cells were measured on day 14. Lines (**a**) and data points (**b–d**) represent individual mice.

virus (LCMV) Armstrong²¹, EMCV or SFV (see Supplementary Fig. 4a, b), or viruses such as influenza that sequester dsRNA²¹, may fail to trigger this pathway. Such viruses can nevertheless be recognized through TLRs specific for viral genomes provided that enough viral particles gain access to the endosomal compartment of DCs, where those receptors are localized²⁴. However, many viruses are only ever 'seen' by DCs in cell-associated form. Here, we show that DCs use TLR3 to detect cell-associated viral dsRNA and that this receptor plays a role *in vivo* in cross-priming against virally infected cells. Our data provide a rationale for the existence of a pattern-recognition receptor for dsRNA that faces into the endosomal compartment, and also explain the apparent restriction of TLR3 expression to those subsets of APCs that avidly phagocytose dying cells^{15,25}. Furthermore, our results offer a possible explanation for the controversial role of TLR3 in antiviral responses. Indeed, TLR3-deficient mice fail to show increased susceptibility to many viral infections²³ and display altered responses to others²⁶. On the basis of our results, one might predict that TLR3-dependence relates directly to the relative contribution of cross-priming versus direct priming to the antiviral response, which is a matter of considerable debate²⁷. Nevertheless, even in infections that strongly depend on cross-priming, the contribution of TLR3 is unlikely to be absolute. Indeed, *Tlr3*^{-/-} mice still mount residual CTL responses to SFV-OVA-infected cells (Fig. 4), clearly indicating the presence of alternative pathways for cross-priming during viral infection.

TLR3 signalling augments cross-presentation by DCs and leads to upregulation of co-stimulatory molecules and production of immunomodulatory cytokines²⁸. All these facets of DC activation are probably responsible for the increased cross-priming response to dsRNA-loaded cells *in vivo*. However, the upregulation of co-stimulatory molecules may be secondary to IFN- α/β induction by TLR3 and autocrine signalling through the IFN- α/β receptor²⁹. Notably, virally infected cells themselves also produce IFN- α/β and activate DCs to promote CTL cross-priming¹³. We have used xenogeneic cells in our experiments in order to eliminate the effect of IFN- α/β derived from the infected cells and to isolate the contribution of TLR3. This may represent the situation that arises during infection with viruses that block IFN- α/β production³⁰. In other viral infections, TLR3 recognition of cell-associated dsRNA, together with IFN- α/β produced by infected cells, are likely to act synergistically to promote DC activation. Thus, in an autologous setting, the cross-priming response to dsRNA-bearing cells is likely to be even more potent than that seen here with xenogeneic cells. Tumour cells loaded with dsRNA could therefore constitute potent vaccines for cancer immunotherapy. □

Methods

Reagents

Poly I:C was obtained from Pharmacia. OVA and latrunculin B were from Calbiochem. Chloroquine was from Sigma. CpG-containing oligonucleotides 1668 and D19 and OVA peptide 257–264 (SIINFEKL) were made at Cancer Research UK. All reagents were free of endotoxin.

Animals and cells

C57BL/6 and C57BL/6 $\times$ 129 F2 mice were obtained from Charles River UK. *Pkr*^{-/-} mice (a gift from C. Weissmann), *Myd88*^{-/-} mice (a gift from S. Akira), OT-I mice on a *Rag1*^{-/-} background, B6.SJL and (B6.SJL $\times$ C57BL/6)F₁ mice were bred at CRUK. *Tlr3*^{-/-} mice¹⁴ (fully backcrossed to C57BL/6) were bred at Yale University or (mixed 129/C57BL/6 background) at the Karolinska Institute. The genetic background had no influence on the results. Bone marrow chimaeras were generated by reconstituting lethally (2 $\times$ 5.5 Gy) gamma-irradiated CD45.1 B6.SJL mice with 2 $\times$ 10⁶ congenic bone marrow cells from either CD45.2 *Tlr3*^{-/-} mice or CD45.2 *Tlr3*^{+/+} littermate controls. All chimaeras were left for at least 6 weeks before use to allow turnover of the splenic DC compartment.

Splenocytes were prepared and subsets of DCs were purified by cell sorting as described¹⁵. CD8 α^+ DCs from chimaeric mice were further sorted on the basis of CD45.2 expression. Subsets were routinely 95–99% pure. T cells were purified from lymph nodes of OT-I $\times$ *Rag1*^{-/-} mice by negative selection. Vero, COS-7, HEK and HeLa cell lines were obtained from the cell production unit at CRUK.

Virus infection, poly I:C- and OVA-loading

EMCV was a gift from I. Kerr. Production of 'suicide' SFV particles containing SFV-OVA

RNA was performed as described previously²⁰. The heterologous OVA gene is driven by the SFV subgenomic promoter and encodes a non-secreted form of OVA. Vero cells or DCs were infected with EMCV or SFV-OVA virus, cells were collected at 2–7 h and were used for experiments or were fixed in paraformaldehyde and the extent of infection monitored by intracellular staining with rabbit anti-EMCV antibody (gift from I. Kerr). For immunization, cells infected with SFV were subsequently treated with trypsin and succinic acid to inactivate any surface-bound viral particles.

Unless otherwise stated, cells were electroporated with 5 mg OVA and/or 10 μ g poly I:C in a 200 μ l volume, as previously described²¹. Mock-treated cells were electroporated in the absence of poly I:C. Vero cells electroporated in this manner retained approximately 11 ng poly I:C per 10⁶ cells as determined by radioactive tracer incorporation (data not shown). To quantify cell-associated OVA (see text), cells were fixed and stained intracellularly with a mouse anti-OVA monoclonal antibody (clone OVA-14, Sigma) or were lysed and the amount of OVA protein determined by standard ELISA using anti-OVA monoclonal antibody as capture and a polyclonal rabbit anti-OVA (Sigma) detection antibody.

In vitro stimulation

Sorted CD8 α^+ DCs were seeded in 96-well plates at 5–50 $\times$ 10⁴ cells per well. Mock-treated, poly I:C-loaded, EMCV-infected cells or SFV-OVA-infected cells were then added at a 4–5:1 ratio. In some experiments, stimulator cells were also exposed to ultraviolet irradiation (254 nm; 15 mJ cm⁻²) to induce death. DCs were analysed by flow cytometry or confocal microscopy at different time points of co-culture; lysates were prepared for reverse transcriptase polymerase chain reaction (RT-PCR) analysis after 3–4 h. Supernatants were assayed for cytokine content by ELISA after overnight culture. All cytokine data shown are means of triplicate cultures $\pm$ 1 s.d.

CTL cross-priming

Mice were immunized by intravenous or intraperitoneal injection of xenogeneic cells (1 $\times$ 10⁶ per mouse) that had been electroporated with OVA $\pm$ poly I:C or infected with SFV-OVA. Mock-treated Vero cells or PBS were used as controls. In some experiments, mice received OT-I cells (1–10 $\times$ 10⁴ per mouse) one day before immunization. To measure OVA-specific CD8⁺ T cell expansion, spleens were collected 6–14 days later and cell suspensions were stained with H-2Kb/SIINFEKL tetramer (ProImmune Ltd) followed by anti-CD8 α and anti-Thy1.2 antibodies. Cells were analysed by flow cytometry after adding CalBRITE beads (BD Biosciences) to determine the absolute number of tetramer⁺ CD8⁺ cells per spleen. OVA-specific cytotoxicity was measured *ex vivo* by IFN- γ ELISPOT or, 5 days after restimulation, by chromium release assay using peptide-pulsed targets. Alternatively, CTL activity was determined by an *in vivo* killing assay. Briefly, mice received a mixture of splenocytes intravenously (10⁷ cells per mouse) that had been pulsed with different concentrations of OVA peptide and labelled with different amounts (3, 0.3 or 0.03 μ M) of CFSE (Molecular Probes). One day later, animals were sacrificed and the frequency of each target cell population was determined. Appropriate CD45 alleles were used to distinguish target from host cells. Specific killing (%) was calculated using the formula (1 – %CFSE_{peptide}/CFSE_{no peptide}) $\times$ 100.

PCR

TLR3 and TLR7 mRNA was measured as described¹⁵. Analysis of cytokine and co-stimulatory gene expression in CD8 α^+ DCs was carried out using a similar protocol and the following primers (forward; reverse): IL-12p40, β -actin as previously described¹⁵; IFN- α (AGGCTCAAGCCATCCCTGT; AGGCACAGGGGCTGTCTTTCTTCT), IFN- β (TTCTGTGCTGTCTTCCAC; GATCACTACCACTCCAGAGCT), TNF- α (GTTCTGCAAAGGAGAGTGG; TGGTCAACCAATCAGCGTGA), IL-6 (GTTCTCTGGGAAATCGTGGA; GTGACTCCAGGTAGCTATGG); hypoxanthine guanine phosphoribosyl transferase (HPRT, GCTGGTGAAAGGACCTCT; CACAGGACTAGAACACCTGC).

Received 29 November 2004; accepted 4 January 2005; doi:10.1038/nature03326.

Published online 13 February 2005.

- Heath, W. R. *et al.* Cross-presentation, dendritic cell subsets, and the generation of immunity to cellular antigens. *Immunol. Rev.* **199**, 9–26 (2004).
- Fonteneau, J. F. *et al.* Activation of influenza virus-specific CD4⁺ and CD8⁺ T cells: a new role for plasmacytoid dendritic cells in adaptive immunity. *Blood* **101**, 3520–3526 (2003).
- Salio, M., Palmowski, M. J., Atzberger, A., Hermans, I. F. & Cerundolo, V. CpG-matured murine plasmacytoid dendritic cells are capable of *in vivo* priming of functional CD8 T cell responses to endogenous but not exogenous antigens. *J. Exp. Med.* **199**, 567–579 (2004).
- Allan, R. S. *et al.* Epidermal viral immunity induced by CD8 α dendritic cells but not by Langerhans cells. *Science* **301**, 1925–1928 (2003).
- Smith, C. M. *et al.* Cutting edge: conventional CD8 α (+) dendritic cells are preferentially involved in CTL priming after footpad infection with herpes simplex virus-1. *J. Immunol.* **170**, 4437–4440 (2003).
- Belz, G. T. *et al.* Cutting edge: conventional CD8 α (+) dendritic cells are generally involved in priming CTL immunity to viruses. *J. Immunol.* **172**, 1996–2000 (2004).
- den Haan, J. M., Lehar, S. M. & Bevan, M. J. CD8(+) but not CD8(-) dendritic cells cross-prime cytotoxic T cells *in vivo*. *J. Exp. Med.* **192**, 1685–1696 (2000).
- Valdez, Y. *et al.* Major histocompatibility complex class II presentation of cell-associated antigen is mediated by CD8 α dendritic cells *in vivo*. *J. Exp. Med.* **195**, 683–694 (2002).
- Iyoda, T. *et al.* The CD8(+) dendritic cell subset selectively endocytoses dying cells in culture and *in vivo*. *J. Exp. Med.* **195**, 1289–1302 (2002).
- Schulz, O. & Reis e Sousa, C. Cross-presentation of cell-associated antigens by CD8 α^+ dendritic cells is attributable to their ability to internalise dead cells. *Immunology* **107**, 183–189 (2002).
- Belz, G. T. *et al.* The CD8 α (+) dendritic cell is responsible for inducing peripheral self-tolerance to tissue-associated antigens. *J. Exp. Med.* **196**, 1099–1104 (2002).
- Liu, K. *et al.* Immune tolerance after delivery of dying cells to dendritic cells *in situ*. *J. Exp. Med.* **196**, 1091–1097 (2002).

13. Le Bon, A. *et al.* Cross-priming of CD8⁺ T cells stimulated by virus-induced type I interferon. *Nature Immunol.* **4**, 1009–1015 (2003).
14. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF- κ B by Toll-like receptor 3. *Nature* **413**, 732–738 (2001).
15. Edwards, A. D. *et al.* Toll-like receptor expression in murine DC subsets: lack of TLR7 expression by CD8 α^+ DC correlates with unresponsiveness to imidazoquinolines. *Eur. J. Immunol.* **33**, 827–833 (2003).
16. Kaufman, R. J. Double-stranded RNA-activated protein kinase mediates virus-induced apoptosis: a new role for an old actor. *Proc. Natl Acad. Sci. USA* **96**, 11693–11695 (1999).
17. Harshyne, L. A., Watkins, S. C., Gambotto, A. & Barratt-Boyes, S. M. Dendritic cells acquire antigens from live cells for cross-presentation to CTL. *J. Immunol.* **166**, 3717–3723 (2001).
18. Helwig, F. & Schmidt, E. A filter-passing agent producing interstitial myocarditis in anthropoid apes and small animals. *Science* **102**, 31–33 (1945).
19. Strauss, J. H. & Strauss, E. G. The alphaviruses: gene expression, replication, and evolution. *Microbiol. Rev.* **58**, 491–562 (1994).
20. Smerdou, C. & Liljestrom, P. Two-helper RNA system for production of recombinant Semliki forest virus particles. *J. Virol.* **73**, 1092–1098 (1999).
21. Diebold, S. S. *et al.* Viral infection switches non-plasmacytoid dendritic cells into high interferon producers. *Nature* **424**, 324–328 (2003).
22. Freigang, S., Egger, D., Bienz, K., Hengartner, H. & Zinkernagel, R. M. Endogenous neosynthesis vs. cross-presentation of viral antigens for cytotoxic T cell priming. *Proc. Natl Acad. Sci. USA* **100**, 13477–13482 (2003).
23. Edelmann, K. H. *et al.* Does Toll-like receptor 3 play a biological role in virus infections? *Virology* **322**, 231–238 (2004).
24. Crozat, K. & Beutler, B. TLR7: A new sensor of viral infection. *Proc. Natl Acad. Sci. USA* **101**, 6835–6836 (2004).
25. Muzio, M. *et al.* Differential expression and regulation of toll-like receptors (TLR) in human leukocytes: selective expression of TLR3 in dendritic cells. *J. Immunol.* **164**, 5998–6004 (2000).
26. Tabet, K. *et al.* Toll-like receptors 9 and 3 as essential components of innate immune defense against mouse cytomegalovirus infection. *Proc. Natl Acad. Sci. USA* **101**, 3516–3521 (2004).
27. Melief, C. J. Regulation of cytotoxic T lymphocyte responses by dendritic cells: peaceful coexistence of cross-priming and direct priming? *Eur. J. Immunol.* **33**, 2645–2654 (2003).
28. Reis e Sousa, C. Toll-like receptors and dendritic cells: for whom the bug tolls. *Semin. Immunol.* **16**, 27–34 (2004).
29. Hoebe, K. & Beutler, B. LPS, dsRNA and the interferon bridge to adaptive immune responses: Trif, Tram, and other TIR adaptor proteins. *J. Endotoxin Res.* **10**, 130–136 (2004).
30. Katze, M. G., He, Y. & Gale, M. Viruses and interferon: A fight for supremacy. *Nature Rev. Immunol.* **2**, 675–687 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was funded by Cancer Research UK (C.R.S.) and by the Swedish Research Council and the EU program (P.L.). We thank I. Kerr for providing EMCV and anti-EMCV antiserum, L. van Dinten for suggestions on 'suicide' virus models and L. Kostic for technical assistance. We are grateful to R. Germain, I. Kerr and members of the Immunobiology Laboratory, Cancer Research UK, for advice and critical review of the manuscript. R.A.F. is an investigator of the Howard Hughes Medical Institute, M.A.N. is supported by an EMBO long-term fellowship and Y.T.A. is supported by the Nakatomi Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.R.S. (caetano@cancer.org.uk).

State transitions and light adaptation require chloroplast thylakoid protein kinase STN7

Stéphane Bellafiore¹, Frédy Barneche¹, Gilles Peltier² & Jean-David Rochaix¹

¹Departments of Molecular Biology and Plant Biology, University of Geneva, 30, Quai Ernest Ansermet, 1211 Geneva, Switzerland

²CEA Cadarache, DSV, DEVM, Laboratoire d'Ecophysiologie de la Photosynthèse, UMR 6191 CNRS-CEA, Aix Marseille II, F-3108 Saint-Paul-Durance, France

Photosynthetic organisms are able to adjust to changing light conditions through state transitions, a process that involves the redistribution of light excitation energy between photosystem II (PSII) and photosystem I (PSI)^{1,2}. Balancing of the light absorption capacity of these two photosystems is achieved through the

reversible association of the major antenna complex (LHCII) between PSII and PSI (ref. 3). Excess stimulation of PSII relative to PSI leads to the reduction of the plastoquinone pool and the activation of a kinase^{4,5}; the phosphorylation of LHCII; and the displacement of LHCII from PSII to PSI (state 2). Oxidation of the plastoquinone pool by excess stimulation of PSI reverses this process (state 1). The *Chlamydomonas* thylakoid-associated Ser-Thr kinase Stt7, which is required for state transitions, has an orthologue named STN7 in *Arabidopsis*⁶. Here we show that loss of STN7 blocks state transitions and LHCII phosphorylation. In *stn7* mutant plants the plastoquinone pool is more reduced and growth is impaired under changing light conditions, indicating that STN7, and probably state transitions, have an important role in response to environmental changes.

Although the phosphorylation of LHCII was observed many years ago^{7,8}, the search for kinases involved in this process in vascular plants has not yet been successful^{9,10}. The *Arabidopsis* genome contains two genes, *STN7* and *STN8*, that display significant sequence identity with the *Chlamydomonas* gene encoding the chloroplast Stt7 protein Ser-Thr kinase⁶. To determine the function of these proteins two *Arabidopsis* lines with T-DNA insertions in these genes were obtained from the Salk Institute collection. After self-crosses, homozygous lines for three T-DNA insertions (Supplementary Fig. S1) were identified by polymerase chain reaction (PCR) on genomic DNA by using appropriate primers as described in the Methods (data not shown). RT-PCR with specific primers for *STN7* and *STN8* was performed with RNA from the wild type and from the *stn7* and *stn8* mutants. Fragments corresponding to *STN7* and *STN8* with the expected size and sequence could be amplified from the wild-type RNA but not from the RNA of the mutants, indicating that the expression of the *STN7* and *STN8* genes is blocked in these lines (Supplementary Fig. S1).

In land plants, 15–20% of LHCII is mobile during state transitions and is reversibly displaced between PSII and PSI (ref. 3). To determine whether state transitions are affected in *stn7*, fluorescence measurements were performed as described^{11,12}. The maximum fluorescence signal, F_m , was measured on an intact leaf with a saturating flash using a pulse amplitude modulation fluorimeter. The leaf was subsequently illuminated with blue light in order to excite preferentially PSII, and the stationary fluorescence yield was recorded. After 15 min, far-red light was added to the blue light. This led to the stimulation of PSI and the transition to state 1 (Fig. 1a, b). After 15 min of blue and far-red light treatment, the maximal fluorescence in state 1 (F_{m1}) was determined. Then the far-red light was switched off to promote the return to state 2 under blue light excitation, and the maximal fluorescence in state 2 (F_{m2}) was determined after 15 min. Because the intensity of the light used to induce state transitions in Fig. 1 was not sufficient to elicit photo-inhibition as verified by F_v/F_m (where F_v is variable fluorescence) measurements (data not shown), we conclude that the observed changes in F_m are caused by state transitions alone.

Transition from state 1 to state 2 can be measured by the changes in maximal fluorescence ($(F_{m1} - F_{m2})/F_{m1}$)100 (ref. 8). In the wild-type strain this value was 10% (Fig. 1a). A similar value was obtained with heterozygous *stn7/stn7* plants (data not shown). In contrast, in the homozygous *stn7/stn7* mutant, state transitions were undetectable (Fig. 1b). Thus the *stn7* mutation is recessive, as expected from a loss-of-function mutation. Similar measurements with the *stn8* homozygous mutant indicated that it is not significantly affected in state transitions (data not shown). The double mutant *stn7/stn8* displayed the same phenotype as *stn7*. The wild-type phenotype was restored after *stn7* plants were transformed with the wild-type STN7 gene (Fig. 1c).

To determine the changes in fluorescence both in PSII and PSI, low-temperature fluorescence emission spectra were measured at 77 K under state 1 and state 2 conditions. The spectra were normalized at 685 nm, corresponding to the peak of PSII fluorescence. In

wild-type plants a transition from state 1 to state 2 was accompanied by a large increase in relative PSI fluorescence at 730 nm, indicating a redistribution of the light excitation energy from PSII to PSI (Fig. 1g). In contrast, only a small increase in relative PSI fluorescence occurred in the *stn7* mutant, confirming that transition to state 2 is vastly reduced and that this mutant is blocked in state 1, similar to the *stt7* mutant of *Chlamydomonas* (Fig. 1h).

The presence of a putative NH₂-terminal transit peptide suggested that the *STN7* kinase is localized within the chloroplast, as is the case with the *Stt7* orthologue in *Chlamydomonas*⁶. To verify this localization, we transformed the *stn7* mutant with a construct carrying the *STN7* complementary DNA tagged with the haemagglutinin (HA) epitope at its 3' end under the control of the endogenous *STN7* promoter. Transformants obtained with this construct had a wild-type state transition phenotype, indicating that the tagged protein retains its activity (Fig. 1d). Immunoblotting of purified soluble and membrane chloroplast fractions with anti-HA antiserum revealed that STN7-HA was present in the chloroplast membrane fraction, with the expected size of 58 kDa (Supplementary Fig. S2A). Treatment of chloroplast membranes with calf intestinal phosphatase increased the mobility of the band, suggesting that the STN7 protein is phosphorylated (Supplementary Fig. S2A). The chloroplast localization of STN7 was further confirmed by transient expression of STN7 fused to green fluorescent protein (GFP) in isolated *Arabidopsis* protoplasts. Moreover, the amino-terminal 88 amino acids of STN7 are sufficient to target GFP into chloroplasts (Supplementary Fig. S2B).

The current model of state transitions assumes that phosphorylation of LHCII is required for the transition from state 1 to state 2 (refs 12, 13). Thr residues in the N-terminal part of the products of *Lhcb1* and *Lhcb2* are phosphorylated under state 2 conditions⁷. If the *stn7* mutant is blocked in state 1, it should fail to phosphorylate

LHCII under state 2 conditions. To test this prediction, LHCII phosphorylation was measured under state 1 and state 2 conditions in the wild-type and *stn7* strains. State 1 was obtained by maintaining the plants in the dark¹⁴ and state 2 was induced by light (80 $\mu\text{mol m}^{-2} \text{s}^{-1}$, 30 min). Thylakoid membranes were isolated and the proteins were fractionated by polyacrylamide gel electrophoresis (PAGE) and immunoblotted with anti-phosphothreonine antibodies. A marked increase in LHCII phosphorylation was observed for the wild type but not for *stn7* under state 2 conditions (Fig. 2a). Moreover, phosphorylation of the D1 and D2 proteins was enhanced under these conditions for both the wild type and *stn7*. The fact that the increase in D1/D2 phosphorylation still occurred in *stn7* under state 2 conditions indicates that the STN7 kinase is specifically involved in LHCII phosphorylation. Protein kinase activity was also determined with isolated thylakoids using [γ -³²P]ATP for labelling. State 1 was obtained in the dark and gave rise to a weak phosphorylation of LHCII. Upon induction of state 2 by light for 20 min, LHCII was strongly phosphorylated in wild-type but not in *stn7* thylakoids (Fig. 2b). Phosphorylation of D1/D2 and PsbH was also markedly increased in wild-type thylakoids and only slightly less so in *stn7* thylakoids, indicating that these different substrates are phosphorylated by another kinase. The identity of the phosphorylated LHCII band was confirmed by its enrichment in the purified LHCII fraction (data not shown).

To test further the role of the STN7 kinase in state transitions, the conserved Lys 167 within domain II of the HA-tagged STN7 kinase was changed to Arg or Gln by site-directed mutagenesis. The gene of this altered kinase was introduced into the *stn7* mutant by transformation. No restoration of state transitions was observed, although the mutant protein accumulated (Fig. 1d–f, i). This clearly demonstrates that the activity of the STN7 kinase is required for state transitions. Thus, the *Chlamydomonas* Stt7 and *Arabidopsis* STN7 kinases are not only structurally but also functionally related. However, it is not yet known whether these kinases phosphorylate LHCII directly or whether additional kinases are involved in this process.

Mutants deficient in cytochrome *b₆* or PSI activity are deficient in state transitions^{15,16}. To test whether accumulation of these complexes was affected in *stn7*, immunoblots were performed with proteins from wild type and the *stn7* mutant using antisera directed against the Rieske and PsaA proteins, which are representative subunits of these complexes. Moreover, immunoblots with antibodies against D1 (PSII), AtpA, LHCII and LHCI were used. In all cases no difference between wild type and *stn7* was observed (Supplementary Fig. S4). Because the TAK kinase also appears to be involved in state transitions^{17,18}, we verified by immunoblotting with TAK antisera that its expression is not altered in the *stn7*

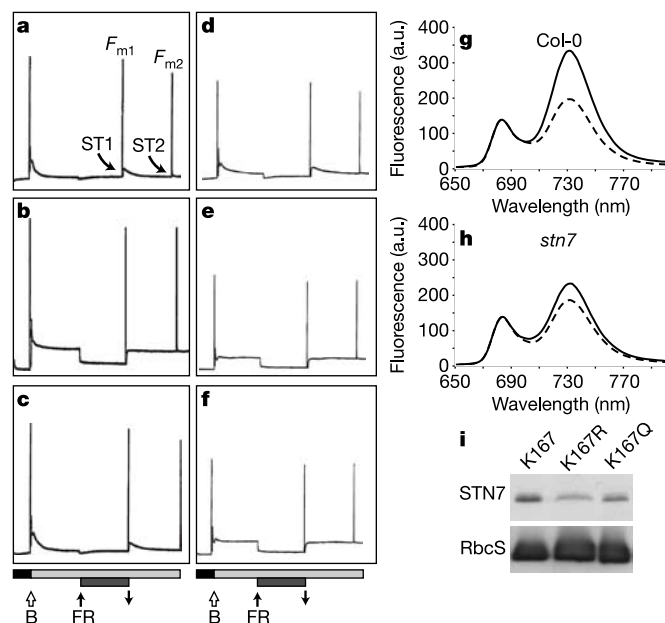


Figure 1 The STN7 protein kinase of *Arabidopsis* is required for state transitions. **a–f**, Blue light (B) and blue light supplemented with far-red light (FR) were used to induce transitions to state 2 and state 1, respectively. F_m (maximal fluorescence) was measured with a saturating 0.8-s flash at room temperature. Upward arrow, light switched on; downward arrow, light switched off. **a**, Col-0 (wild type); **b**, *stn7*; **c**, *stn7*-R, *stn7* rescued with STN7; **d–f**, *stn7* transformed with STN7-HA, STN7-HA-K167R and STN7-HA-K167Q, respectively. **g**, **h**, Low-temperature emission spectra of thylakoids from Col-0 (**g**) and *stn7* (**h**) in state 1 (dashed lines) and state 2 (solid lines). a.u., arbitrary units. **i**, Immunoblots with total extracts from *stn7* transformed with STN7-HA, STN7-HA-K167R and STN7-HA-K167Q with anti-HA and RbcS antisera.

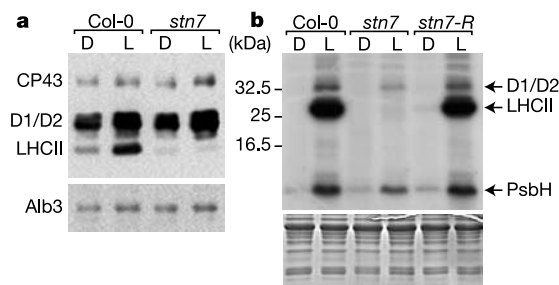


Figure 2 Phosphorylation of LHCII is diminished in *stn7* under state 2 conditions. **a**, Thylakoid membrane proteins extracted from Col-0 and *stn7* in the dark (state 1) or in the light (state 2) were separated by PAGE and immunoblotted with an anti-phosphothreonine antiserum. Equal protein loading was checked with antiserum against the thylakoid protein Alb3. **b**, Thylakoid membrane proteins extracted from wild-type and *stn7* leaves were incubated with [γ -³²P]ATP in the dark (state 1) or in the light (state 2). Proteins were separated by PAGE and autoradiographed.

mutant (Supplementary Fig. S4). On the basis of the observation that loss of the TAK and STN7 kinases leads to different phenotypes and that in contrast to STN7, the TAK kinase is involved in the phosphorylation of other thylakoid proteins besides LHCII, it is unlikely that these kinases act in the same pathway.

Because the main component in non-photochemical chlorophyll fluorescence quenching is energy-dependent (qE)¹⁹, it was important to check that the observed changes in fluorescence of the *stn7* mutant are not due to alterations in qE. Wild-type and *stn7* plants were grown either at a PFD (photon flux density) of 60 or 160 $\mu\text{mol m}^{-2} \text{s}^{-1}$ and detached leaves were subsequently challenged with increasing light intensities. There was no significant difference in non-photochemical chlorophyll fluorescence quenching (NPQ) between wild-type and mutant plants under both conditions (Fig. 3a, b). Moreover, we confirmed previous studies showing that the qE-deficient mutants are able to perform state transitions²⁰ (data not shown). Taken together these results indicate that state transitions and qE operate independently from each other.

Although state transitions were discovered more than thirty years ago^{1,2}, their exact function is still not clear. The proposal that this process could have a role in protection against high light levels seems less likely given the fact that the LHCII kinase is inactivated by high light levels²¹. Another proposal is that state transitions may be involved in optimizing the photosynthetic yield and thus growth under low light conditions. However, growth of *stn7* mutant plants was not different compared to that of wild type when plants were subjected to an 8 h light/16 h dark regime either at a low (60 $\mu\text{mol m}^{-2} \text{s}^{-1}$) or higher PFD (160 $\mu\text{mol m}^{-2} \text{s}^{-1}$). Photosynthetic parameters were determined from detached leaves of these plants by measuring chlorophyll fluorescence and CO₂ assimilation under increasing light intensities. In *stn7* leaves, the plastoquinone pool was more reduced as shown by the increase in $1 - q_P$ (Fig. 3c, d), where q_P is photochemical quenching. This is probably

due to the fact that excitation energy could not be redistributed to PSI through state transitions. An increase in plastoquinone reduction was also reported in the *Arabidopsis* strain lacking PsaH, which is deficient in state transitions^{22,23}. Although CO₂ assimilation rates of *stn7* and wild-type leaves were similar in plants grown at 160 $\mu\text{mol m}^{-2} \text{s}^{-1}$, this was no longer the case when plants were grown at 60 $\mu\text{mol m}^{-2} \text{s}^{-1}$ (Fig. 3e, f). Under these conditions the maximum rate of CO₂ fixation (P_{max}) was reduced but there was no significant change of the quantum yield of CO₂ fixation (initial slope in Fig. 3e). These observations are compatible with the fact that there was no significant change in growth between wild type and *stn7* under low light conditions. The lowering of P_{max} may result from the loss of control of the plastoquinone redox state in *stn7*, which may perturb the adaptation processes to light, which are known to be complex. Thus, the STN7 kinase may also have an important role in the regulation of the redox state of the plastoquinone pool by balancing the light absorption between the two photosystems.

In order to determine the impact of state transitions on plant growth, plants were placed under changing light conditions. In a first experiment, plants were grown under 8 h light/16 h dark photoperiods in which the PFD was changed every hour from 50 to 240 $\mu\text{mol m}^{-2} \text{s}^{-1}$ during the light period. In another experiment the plants were subjected to a 12 h light/12 h dark regime with alternating 1 h PSII and PSI light cycles during the light phase (see Methods). Under both conditions growth of the *stn7* mutant was impaired relative to wild-type plants as further confirmed by fresh and dry weight measurements (Fig. 4). Notably, flowering of the *stn7* plants grown under changing PSII and PSI light cycles occurred earlier (44 ± 3 days) than in the wild-type plants (51 ± 4 days). When the periods of the light cycles were increased, the differences in growth between wild-type and mutant plants became less pronounced. It is likely that under these conditions other more long-term mechanisms of adaptation involving changes in the expression of photosynthetic genes²⁴ become prominent. Our results differ from those obtained with *Arabidopsis* plants that lack PsaH and have state transition deficiencies. In this case growth was not impaired although oxygen evolution was diminished by 14%²³. These differences could be due to the particular growth conditions used and/or to the different primary lesions in these plants.

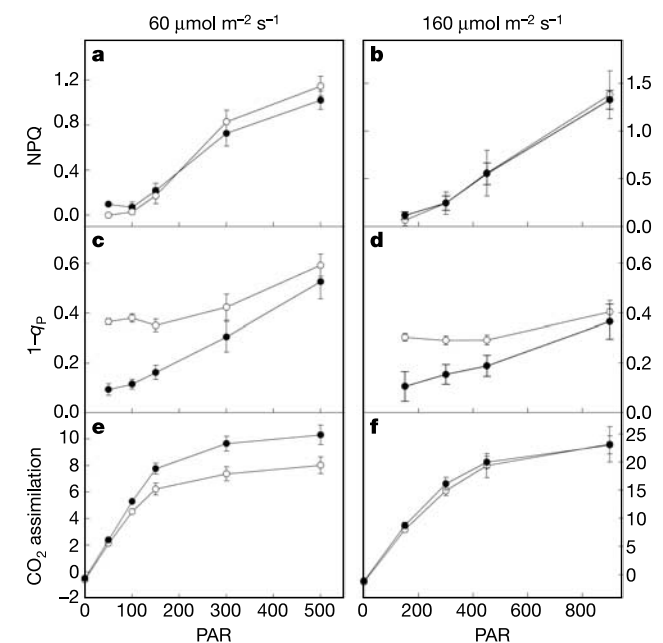


Figure 3 Measurements of photosynthetic parameters for *stn7* detached leaves. Col-0 and *stn7* were grown under an 8 h/16 h light/dark regime at 60 or 160 $\mu\text{mol m}^{-2} \text{s}^{-1}$ during the light period. Measurements of chlorophyll fluorescence parameters and CO₂ assimilation were performed with various light intensities as indicated. Gas conditions were 1.2% O₂, 750 $\mu\text{l CO}_2 \text{l}^{-1}$ air. **a, b**, NPQ (qE); **c, d**, $1 - q_P$; **e, f**, CO₂ assimilation (expressed in $\mu\text{mol CO}_2 \text{m}^{-2} \text{s}^{-1}$). Col-0, filled circles; *stn7*, open circles; PAR, photosynthetic active radiation ($\mu\text{mol m}^{-2} \text{s}^{-1}$). Standard deviations were determined from five and six independent measurements for plants adapted to 60 and 160 $\mu\text{mol m}^{-2} \text{s}^{-1}$, respectively.

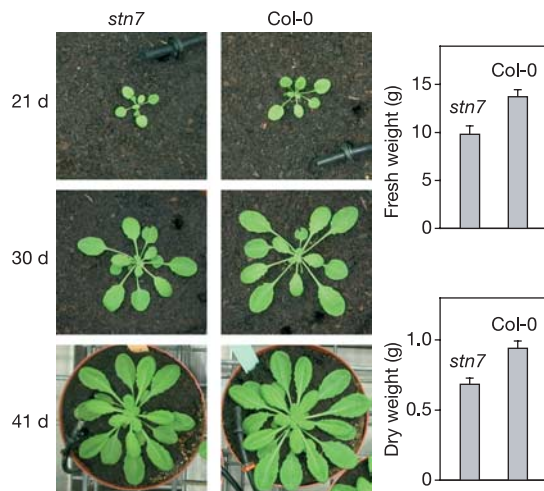


Figure 4 Growth of the *stn7* mutant is impaired under changing light conditions. Col-0 and *stn7* plants were grown under an 8 h/16 h light/dark regime. After one week at 160 $\mu\text{mol m}^{-2} \text{s}^{-1}$, plants were alternatively illuminated for 1 h at 50 and 1 h at 240 $\mu\text{mol m}^{-2} \text{s}^{-1}$. Pictures were taken after 21, 30 and 41 days. Plant fresh and dry weights were determined after 50 days. Standard deviations were determined from eight and seven measurements for *stn7* and Col-0, respectively.

As much as 80% of the LHC antenna is mobile in *Chlamydomonas* during state transitions²⁵, whereas in land plants the mobile fraction of LHCII is only 15–20%³. In spite of this relatively low value, our study shows that in land plants the STN7 kinase, and probably state transitions, are important for adaptation and that in their absence growth is significantly impaired under conditions in which light quality and quantity change frequently. This points to the importance of state transitions in a natural environment where plants are often subjected to light fluctuations of this sort. □

Methods

Plant material

Arabidopsis thaliana (L.) ecotype Columbia (Col-0) was used for all experiments. Plants were grown under controlled conditions of light (50 or 160 $\mu\text{mol m}^{-2} \text{s}^{-1}$; 8 h or 12 h photoperiods, 23/20 °C day/night, and relative air humidity of 50–70%).

All physiological and biochemical analyses were performed with rosette leaves harvested before flowering. We obtained the T-DNA insertion lines in the Columbia background for Atlg68830 (SALK 073254) and At5g01920 (SALK 060869 and SALK 064913) from the Salk Institute (see Supplementary Information for the characterization of these lines and for the DNA, RNA, protein and chlorophyll analyses).

State transitions and NPQ

State transitions and NPQ were measured as described^{11,26} (see Supplementary Information for details).

Photosynthetic measurements

Photosynthetic gas exchange and chlorophyll fluorescence measurements were simultaneously performed on detached leaves using a LI-6400 portable photosynthesis system equipped with a 6400-40 fluorometer (LI-COR Biosciences) (for details see Supplementary Information). PSII light was obtained with cool white fluorescent lamps (Osram L18W/20) with orange 105 Lee filters and PSI light was obtained with red fluorescent lamps (Osram L18W/60) with red 027 Lee filters. Chlorophyll fluorescence emission spectra of thylakoid membrane suspensions were recorded in liquid nitrogen (77 K) as described^{27,28} (see Supplementary Information).

In vivo and in vitro phosphorylation of the LHCII antennae

Leaves from dark-acclimated plants, floating on water, were exposed to low light (80 $\mu\text{mol m}^{-2} \text{s}^{-1}$) or kept in the dark for 30 min²⁹. Thylakoid membranes were isolated from the dark-incubated and illuminated leaves as described²⁷ in the presence of 10 mM NaF to inhibit phospho-LHCII phosphatase activity. Thylakoids were re-suspended in assay buffer consisting of 50 mM HEPES-KOH pH 7.5, 100 mM sucrose, 5 mM NaCl, 10 mM MgCl_2 and 10 mM NaF at a final chlorophyll concentration of 0.4 mg ml^{-1} .

After dark adaptation, thylakoids were isolated from plants according to ref. 30, and re-suspended in storage buffer (100 mM sorbitol, 5 mM MgCl_2 , 5 mM NaCl and 50 mM HEPES/KOH pH 7.5). They were used as substrate for the kinase assay. Thylakoid membrane proteins equivalent to 8 μg of chlorophyll were subjected to a 20 min light induction (80 $\mu\text{mol m}^{-2} \text{s}^{-1}$) at 25 °C in the presence of 10 μCi [γ -³²P]ATP (Amersham 3,000 Ci mmol^{-1}), 0.4 mM ATP and 10 mM NaF in 100 μl of storage buffer²⁹. Reactions were terminated by centrifugation, washing twice in storage buffer and addition of denaturing sample buffer, and were electrophoresed on 12% polyacrylamide-SDS gels, and finally analysed with a phosphorimager.

Received 23 September; accepted 17 December 2004; doi:10.1038/nature03286.

1. Bonaventura, C. & Myers, J. Fluorescence and oxygen evolution from *Chlorella pyrenoidosa*. *Biochim. Biophys. Acta* **189**, 366–383 (1969).
2. Murata, N. Control of excitation transfer in photosynthesis. I. Light-induced change of chlorophyll a fluorescence in *Porphyridium cruentum*. *Biochim. Biophys. Acta* **172**, 242–251 (1969).
3. Allen, J. F. Protein phosphorylation in regulation of photosynthesis. *Biochim. Biophys. Acta* **1098**, 275–335 (1992).
4. Vener, A. V., van Kan, P. J., Rich, P. R., Ohad, I. I. & Andersson, B. Plastoquinol at the quinol oxidation site of reduced cytochrome *b_f* mediates signal transduction between light and protein phosphorylation: Thylakoid protein kinase deactivation by a single-turnover flash. *Proc. Natl Acad. Sci. USA* **94**, 1585–1590 (1997).
5. Zito, F. *et al.* The Qo site of cytochrome *b₆f* complexes controls the activation of the LHCII kinase. *EMBO J.* **18**, 2961–2969 (1999).
6. Depege, N., Bellafiore, S. & Rochaix, J. D. Role of chloroplast protein kinase Stt7 in LHCII phosphorylation and state transition in *Chlamydomonas*. *Science* **299**, 1572–1575 (2003).
7. Bennett, J. Phosphorylation of chloroplast membrane polypeptides. *Nature* **269**, 344–346 (1977).
8. Bennett, J. Chloroplast phosphoproteins. Phosphorylation of polypeptides of the light-harvesting chlorophyll protein complex. *Eur. J. Biochem.* **99**, 133–137 (1979).
9. Race, H. L. & Hind, G. A protein kinase in the core of photosystem II. *Biochemistry* **35**, 13006–13010 (1996).
10. Sokolenko, A. *et al.* The 64 kDa polypeptide of spinach may not be the LHCII kinase, but a lumen-located polyphenol oxidase. *FEBS Lett.* **371**, 176–180 (1995).
11. Jensen, P. E., Gilpin, M., Knoetzel, J. & Scheller, H. V. The PSI-K subunit of photosystem I is involved in the interaction between light-harvesting complex I and the photosystem I reaction center core. *J. Biol. Chem.* **275**, 24701–24708 (2000).
12. Allen, J. F., Bennett, J., Steinback, K. E. & Arntzen, C. J. Chloroplast protein phosphorylation couples plastoquinone redox state to distribution of excitation energy between photosystems. *Nature* **291**, 25–29 (1981).

13. Wollman, F. A. & Delepleire, P. Correlation between changes in light energy distribution and changes in thylakoid membrane polypeptide phosphorylation in *Chlamydomonas reinhardtii*. *J. Cell Biol.* **98**, 1–7 (1984).
14. Bassi, R., Giacometti, G. M. & Simpson, D. J. Changes in the organization of stroma membranes induced by *in vivo* state 1-state 2 transition. *Biochim. Biophys. Acta* **935**, 152–165 (1988).
15. Wollman, F. A. & Lemaire, C. Studies on kinase-controlled state transitions in photosystem II and *b₆f* mutants from *Chlamydomonas reinhardtii* which lack quinone-binding proteins. *Biochim. Biophys. Acta* **933**, 85–94 (1988).
16. Zhang, S. & Scheller, H. V. Light-harvesting complex II binds to several small subunits of photosystem I. *J. Biol. Chem.* **279**, 3180–3187 (2004).
17. Snyders, S. & Kohorn, B. D. TAKs, thylakoid membrane protein kinases associated with energy transduction. *J. Biol. Chem.* **274**, 9137–9140 (1999).
18. Snyders, S. & Kohorn, B. D. Disruption of thylakoid-associated kinase 1 leads to alteration of light harvesting in *Arabidopsis*. *J. Biol. Chem.* **276**, 32169–32176 (2001).
19. Baroli, I. & Niyogi, K. K. Molecular genetics of xanthophyll-dependent photoprotection in green algae and plants. *Phil. Trans. R. Soc. Lond. B* **355**, 1385–1394 (2000).
20. Vink, M. *et al.* Light-modulated exposure of the light-harvesting complex II (LHCII) to protein kinase(s) and state transition in *Chlamydomonas reinhardtii* xanthophyll mutants. *Biochemistry* **43**, 7824–7833 (2004).
21. Rintamaki, E., Martinsuo, P., Pursiheimo, S. & Aro, E. M. Cooperative regulation of light-harvesting complex II phosphorylation via the plastoquinol and ferredoxin-thioredoxin system in chloroplasts. *Proc. Natl Acad. Sci. USA* **97**, 11644–11649 (2000).
22. Lunde, C., Jensen, P. E., Haldrup, A., Knoetzel, J. & Scheller, H. V. The PSI-H subunit of photosystem I is essential for state transitions in plant photosynthesis. *Nature* **408**, 613–615 (2000).
23. Lunde, C. *et al.* Plants impaired in state transitions can to a large degree compensate for their defect. *Plant Cell Physiol.* **44**, 44–54 (2003).
24. Pfannschmidt, T., Nilsson, A. & Allen, J. F. Photosynthetic control of chloroplast gene expression. *Nature* **397**, 625–628 (1999).
25. Delosme, R., Olive, J. & Wollman, F. A. Changes in light energy distribution upon state transitions: an *in vivo* photoacoustic study of the wild type and photosynthesis mutants from *Chlamydomonas reinhardtii*. *Biochim. Biophys. Acta* **1273**, 150–158 (1996).
26. Li, X. P., Gilmore, A. M. & Niyogi, K. K. Molecular and global time-resolved analysis of a *psbS* gene dosage effect on pH- and xanthophyll cycle-dependent nonphotochemical quenching in photosystem II. *J. Biol. Chem.* **277**, 33590–33597 (2002).
27. Robinson, H. H. & Yocum, C. F. Cyclic photophosphorylation reactions catalyzed by ferredoxin, methyl viologen and anthraquinone sulfonate. Use of photochemical reactions to optimize redox poisoning. *Biochim. Biophys. Acta* **590**, 97–106 (1980).
28. Weis, E. Chlorophyll fluorescence at 77K in intact leaves: Characterization of a technique to eliminate artifacts related to self-absorption. *Photosynth. Res.* **6**, 73–86 (1985).
29. Zer, H. *et al.* Light affects the accessibility of the thylakoid light harvesting complex II (LHCII) phosphorylation site to the membrane protein kinase(s). *Biochemistry* **42**, 728–738 (2003).
30. Havaux, M., Dall'Osto, L., Cuine, S., Giuliano, G. & Bassi, R. The effect of zeaxanthin as the only xanthophyll on the structure and function of the photosynthetic apparatus in *Arabidopsis thaliana*. *J. Biol. Chem.* **279**, 13878–13888 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank N. Roggli for drawings; C. Niyogi and M. Havaux for the *npq4* mutant; C. Fankhauser for transformation vectors and help with *Arabidopsis*; C. Bréhélin and F. Kessler (Plant Survival NCCR) for help with the protoplast transformation experiments; B. Genty and M. Goldschmidt-Clermont for discussions; M. Péan, A. Beyly and the GRAP team (CEA Cadarache) for support in growing plants under controlled conditions; and B. Delessert for assistance in the phytotron. F.B. was supported by a long-term EMBO fellowship. This work was supported by a grant from the Swiss National Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.-D.R. (jean-david.rochaix@molbio.unige.ch).

Functional cartography of complex metabolic networks

Roger Guimerà & Luis A. Nunes Amaral

NICO and Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA

High-throughput techniques are leading to an explosive growth in the size of biological databases and creating the opportunity to revolutionize our understanding of life and disease. Interpretation of these data remains, however, a major scientific challenge. Here, we propose a methodology that enables us to extract

and display information contained in complex networks^{1–3}. Specifically, we demonstrate that we can find functional modules^{4,5} in complex networks, and classify nodes into universal roles according to their pattern of intra- and inter-module connections. The method thus yields a ‘cartographic representation’ of complex networks. Metabolic networks^{6–8} are among the most challenging biological networks and, arguably, the ones with most potential for immediate applicability⁹. We use our method to analyse the metabolic networks of twelve organisms from three different superkingdoms. We find that, typically, 80% of the nodes are only connected to other nodes within their respective modules, and that nodes with different roles are affected by different evolutionary constraints and pressures. Remarkably, we find that metabolites that participate in only a few reactions but that connect different modules are more conserved than hubs whose links are mostly within a single module.

If we are to extract the significant information from the topology of a large, complex network, knowledge of the role of each node is of crucial importance. A cartographic analogy is helpful to illustrate this point. Consider the network formed by all cities and towns in a country (the nodes) and all the roads that connect them (the links). It is clear that a map in which each city and town is represented by a circle of fixed size and each road is represented by a line of fixed width is hardly useful. Rather, real maps emphasize capitals and important communication lines so that we can obtain scale-specific information at a glance. Similarly, it is difficult, if not impossible, to obtain information from a network with hundreds or thousands of nodes and links, unless the information about nodes and links is conveniently summarized. This is particularly true for biological networks.

Here, we propose a methodology, which is based on the connectivity of the nodes, that yields a cartographic representation of a complex network. The first step in our method is to identify the functional modules^{4,5} in the network. In the cartographic picture, modules are analogous to countries or regions, and enable a coarse-grained, and thus simplified, description of the network. Then we classify the nodes in the network into a small number of system-independent ‘universal roles’.

It is common that social networks have communities of highly interconnected nodes that are less connected to nodes in other communities. Such modular structures have been reported not only in social networks^{5,10–12}, but also in food webs¹³ and biochemical networks^{4,14–16}. It is widely believed that the modular structure of complex networks plays a critical role in their functionality^{4,14,16}. There is therefore a clear need to develop algorithms to identify modules accurately^{5,11,17–20}.

We identify modules by maximizing the network’s modularity^{11,18,21} using simulated annealing²² (see Methods). Simulated annealing enables us to perform an exhaustive search and to minimize the problem of finding sub-optimal partitions. It is noteworthy that, in our method, we do not need to specify a priori the number of modules; rather, this number is an outcome of the algorithm. Our algorithm is able to reliably identify modules in a network whose nodes have as many as 50% of their connections outside their own module (Fig. 1).

When considering modular networks, it is plausible to surmise that the nodes in a network are connected according to the role they fulfil. This fact has been long recognized in the analysis of social networks²³. For example, in a classical hierarchical organization, the chief executive is not directly connected to plant employees but is connected to the members of the board of directors. Such a

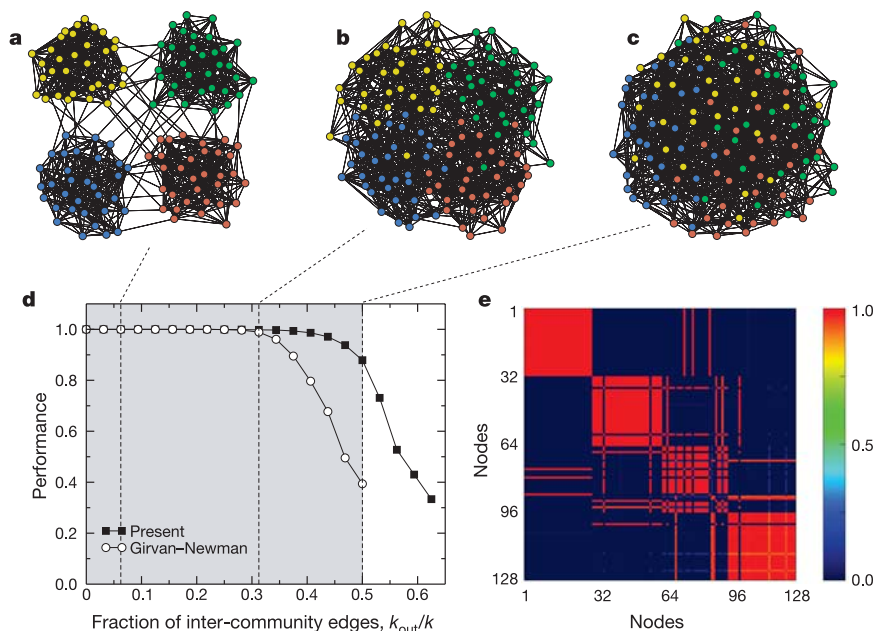


Figure 1 Performance of module identification methods. To test the performance of the method, we build ‘random networks’ with known module structure. Each test network comprises 128 nodes divided into 4 modules of 32 nodes. Each node is connected to the other nodes in its module with probability p_i , and to nodes in other modules with probability $p_o < p_i$. On average, thus, each node is connected to $k_{out} = 96 p_o$ nodes in other modules and to $k_{in} = 31 p_i$ in the same module. Additionally, p_i and p_o are selected so that the average degree of the nodes is $k = 16$. We display networks with: **a**, $k_{in} = 15$ and $k_{out} = 11$; **b**, $k_{in} = 11$ and $k_{out} = 5$; and **c**, $k_{in} = k_{out} = 8$. **d**, The performance of a module identification algorithm is typically defined as the fraction of correctly classified

nodes. We compare our algorithm to the Girvan–Newman algorithm^{5,18}, which is the reference algorithm for module identification^{11,18,19}. Note that our method is 90% accurate even when half of a node’s links are to nodes in outside modules. **e**, Our module-identification algorithm is stochastic, so different runs yield, in principle, different partitions. To test the robustness of the algorithm, we obtain 100 partitions of the network depicted in **c** and plot, for each pair of nodes in the network, the fraction of times that they are classified in the same module. As shown in the figure, most pairs of nodes are either always classified in the same module (red) or never classified in the same module (dark blue), which indicates that the solution is robust.

statement holds for virtually any organization; that is, the role of chief executive is defined irrespective of the particular organization considered.

We propose a new method to determine the role of a node in a complex network. Our approach is based on the idea that nodes with the same role should have similar topological properties²⁴ (see Supplementary Information for a discussion on how our approach relates to previous work). We predict that the role of a node can be determined, to a great extent, by its within-module degree and its participation coefficient, which define how the node is positioned in its own module and with respect to other modules^{25,26} (see Methods). These two properties are easily computed once the modules of a network are known.

The within-module degree z_i measures how ‘well-connected’ node i is to other nodes in the module. High values of z_i indicate high within-module degrees and vice versa. The participation coefficient P_i measures how ‘well-distributed’ the links of node i are among different modules. The participation coefficient P_i is close to 1 if its links are uniformly distributed among all the modules, and 0 if all its links are within its own module.

We define heuristically seven different universal roles, each defined by a different region in the z – P parameter space (Fig. 2). According to the within-module degree, we classify nodes with $z \geq 2.5$ as module hubs and nodes with $z < 2.5$ as non-hubs. Both hub and non-hub nodes are then more finely characterized by using the values of the participation coefficient (see Supplementary Information for a detailed justification of this classification scheme, and for a discussion on possible alternatives).

We find that non-hub nodes can be naturally divided into four different roles: (R1) ultra-peripheral nodes; that is, nodes with all their links within their module ($P \leq 0.05$); (R2) peripheral nodes; that is, nodes with most links within their module ($0.05 < P \leq 0.62$); (R3) non-hub connector nodes; that is, nodes with many links to other modules ($0.62 < P \leq 0.80$); and (R4) non-hub kinless nodes; that is, nodes with links homogeneously distributed among all modules ($P > 0.80$). We find that hub nodes can be naturally divided into three different roles: (R5) provincial hubs; that is, hub nodes with the vast majority of links within their module ($P \leq 0.30$); (R6) connector hubs; that is, hubs with many links to most of the other modules ($0.30 < P \leq 0.75$); and (R7) kinless hubs; that is, hubs with links homogeneously distributed among all modules ($P > 0.75$).

To test the applicability of our approach to complex biological networks, we consider the cartographic representation of the metabolic networks^{6–9,14} of twelve organisms: four bacteria (*Escherichia coli*, *Bacillus subtilis*, *Lactococcus lactis* and *Thermosynechococcus elongatus*), four eukaryotes (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Plasmodium falciparum* and *Homo sapiens*), and four archaea (*Pyrococcus furiosus*, *Aeropyrum pernix*, *Archaeoglobus fulgidus* and *Sulfolobus solfataricus*). In metabolic networks, nodes represent metabolites and two nodes i and j are connected by a link if there is a chemical reaction in which i is a substrate and j a product, or vice versa. In our analysis, we use the database developed by Ma and Zeng⁸ (MZ) from the Kyoto Encyclopedia of Genes and Genomes²⁷ (KEGG). The results we report are not altered if we consider the complete KEGG database instead (Figs 2c and 4b, and Supplementary Information).

First, we identify the functional modules in the different metabolic networks (Fig. 3). Finding modules in metabolic networks purely on the basis of topological properties is an extremely important task. For example, Schuster *et al.* have reported on the impossibility of obtaining elementary flux modes²⁸ from complete metabolic networks due to the combinatorial explosion of the number of such modes²⁹. Our algorithm identifies an average of 15 different modules in each metabolic network—with a maximum of 19 for *E. coli* and *H. sapiens*, and a minimum of 11 for *A. fulgidus*. As expected, the density of links within each of the modules is

significantly larger than between modules—typically 100–1,000 times larger (see Supplementary Information).

To assess how each of the modules is related to the pathways traditionally defined in biology, we use the classification scheme proposed in KEGG, which includes nine major pathways: carbohydrate metabolism, energy metabolism, lipid metabolism, nucleotide metabolism, amino-acid metabolism, glycan biosynthesis and metabolism, metabolism of cofactors and vitamins, biosynthesis of secondary metabolites and biodegradation of xenobiotics. Each metabolite in the KEGG database is assigned to at least one pathway; thus, we can determine to which pathways the metabolites in a given

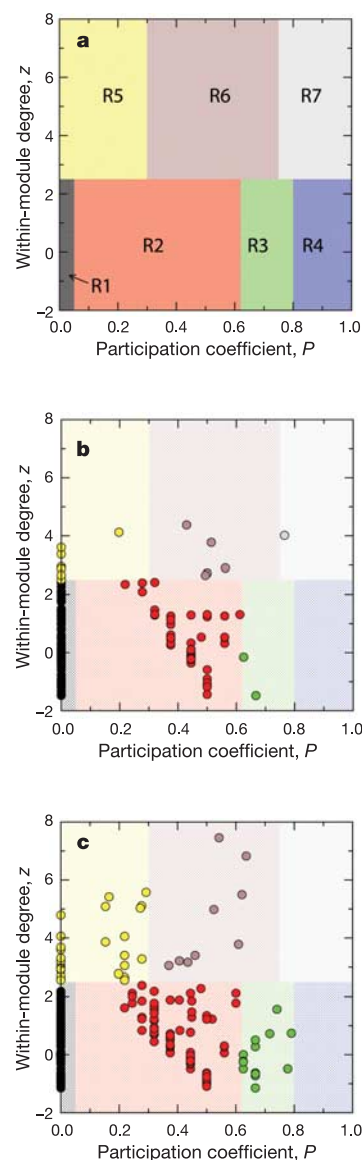


Figure 2 Roles and regions in the z – P parameter space. **a**, Each node in a network can be characterized by its within-module degree and its participation coefficient (see Methods for definitions). We classify nodes with $z \geq 2.5$ as module hubs and nodes with $z < 2.5$ as non-hubs. We find that non-hub nodes can be naturally assigned into four different roles: (R1) ultra-peripheral nodes; (R2) peripheral nodes; (R3) non-hub connector nodes; and (R4) non-hub kinless nodes. We find that hub nodes can be naturally assigned into three different roles: (R5) provincial hubs; (R6) connector hubs; and (R7) kinless hubs (see text and Supplementary Information for details). **b**, Metabolite role determination for the metabolic network of *E. coli*, as obtained from the MZ database. Each metabolite is represented as a point in the z – P parameter space, and is coloured according to its role. **c**, Same as **b** but for the complete KEGG database.

module belong. We find that most modules contain metabolites mostly from one major pathway. For example, in 17 of the 19 modules identified for *E. coli*, more than one-third of the metabolites belong to a single pathway. Interestingly, some other modules—two in the case of *E. coli*—cannot be trivially associated with a single traditional pathway. These modules are typically central in the metabolism and contain, mostly, metabolites that are classified in KEGG as belonging to carbohydrate and amino-acid metabolism.

Next we identify the role of each metabolite. In Fig. 2b we show the roles identified in the metabolic network of *E. coli*. Other organisms show a similar distribution of the nodes in the different roles, even though they correspond to organisms that are very distant from an evolutionary standpoint (see Supplementary Information). Role R1, which contains ultra-peripheral metabolites with small degree and no between-module links, comprises 76–86% of all the metabolites in the networks. This considerably simplifies the coarse-grained representation of the network as these nodes do not need to be identified separately. Note that this finding alone represents an important step towards the goal of extracting scale-specific information from complex networks.

The information about modules and roles enables us to build a cartographic representation of the metabolic network of, for example, *E. coli* (Fig. 3). This representation enables us to recover relevant biological information. For instance, we find that the metabolism is mostly organized around the module containing pyruvate, which in turn is strongly connected to the module whose hub is acetyl-coenzyme A (CoA). These two molecules are key to connecting the metabolism of carbohydrates, amino acids and lipids to the tricarboxylic acid (TCA) cycle from which ATP is obtained.

These two modules are connected to more peripheral ones by key metabolites such as D-glyceraldehyde 3-phosphate and D-fructose 6-phosphate (which connect to the glucose and galactose metabolisms), D-ribose 5-phosphate (which connects to the metabolism of certain nucleotides), and glycerone phosphate (which connects to the metabolism of certain lipids).

Our analysis also uncovers nodes with key connector roles that take part in only a small but fundamental set of reactions. For example, *N*-carbamoyl-L-aspartate takes part in only three reactions but is vital because it connects the pyrimidine metabolism, whose hub is uracil, to the core of the metabolism through the alanine and aspartate metabolism. The potential importance of such non-hub connectors points to another consideration. It is a plausible hypothesis that nodes with different roles are under different evolutionary constraints and pressures. In particular, we expect that nodes with structurally relevant roles are more necessary and therefore more conserved across species.

To quantify the relation between roles and conservation, we define the loss rate $p_{\text{lost}}(R)$ (see Methods). Structurally relevant roles are expected to have low values of $p_{\text{lost}}(R)$ and vice versa. We find that the different roles have different loss rates (Fig. 4). As expected, ultra-peripheral nodes (role R1) have the highest loss rate whereas connector hubs (role R6) are the most conserved across all species considered.

The results for the comparison of $p_{\text{lost}}(R)$ for ultra-peripheral nodes and connector hubs is illustrative, but hardly surprising. The comparison of $p_{\text{lost}}(R)$ for non-hub connectors (role R3) and provincial hubs (role R5), however, yields a surprising finding. The metabolites in the provincial hubs class have many within-module

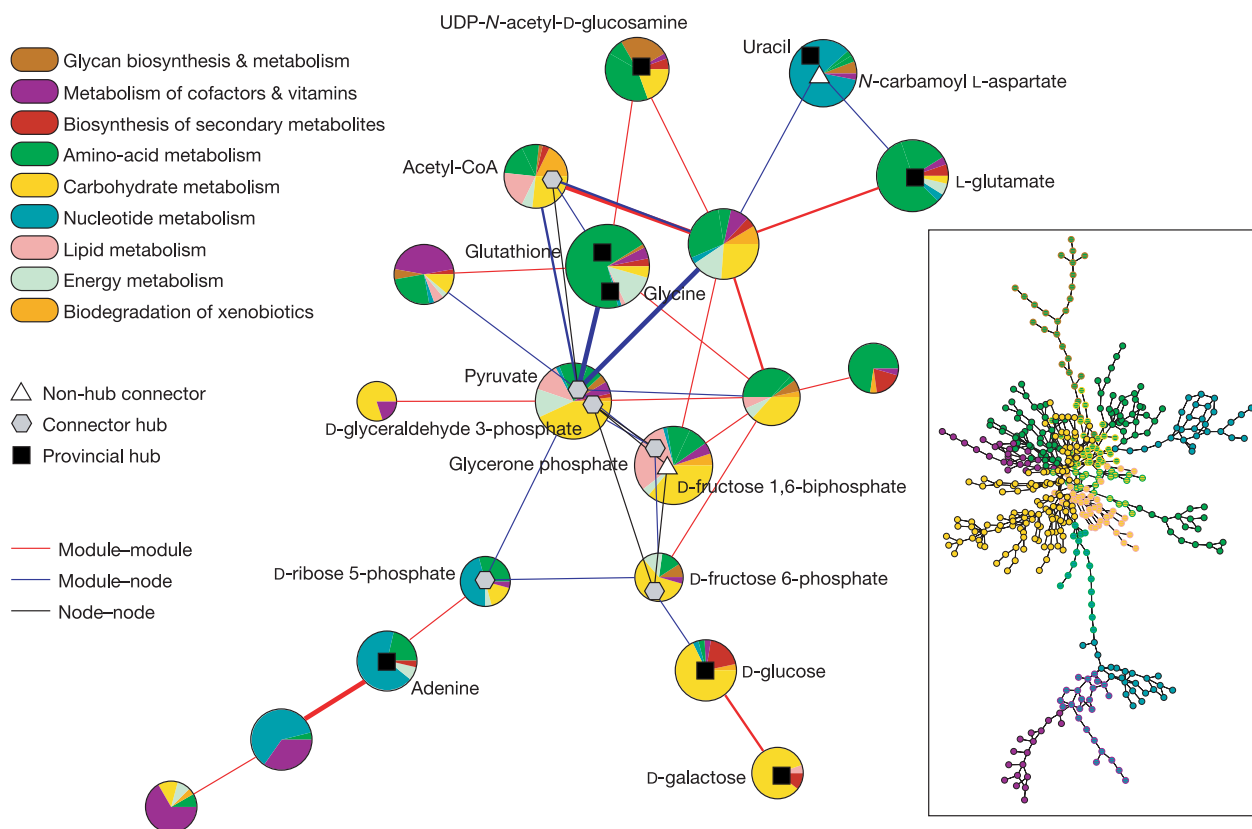


Figure 3 Cartographic representation of the metabolic network of *E. coli*. Each circle represents a module and is coloured according to the KEGG pathway classification of the metabolites it contains. Certain important nodes are depicted as triangles (non-hub connectors), hexagons (connector hubs) and squares (provincial hubs). Interactions between modules and nodes are depicted using lines, with thickness proportional to the

number of actual links. Inset: metabolic network of *E. coli*, which contains 473 metabolites and 574 links. This representation was obtained using the program Pajek. Each node is coloured according to the 'main' colour of its module, as obtained from the cartographic representation.

connections, sometimes as many as five standard deviations more connections than the average node in the module. Conversely, non-hub connector metabolites have few links relative to other nodes in their modules—and fewer total connections than the metabolites in role R5 (see Supplementary Fig. S12b, c). The links of non-hub connectors, however, are distributed among several different

modules, whereas the links of provincial hubs are mainly within their modules. We find that non-hub connectors are systematically and significantly more conserved than provincial hub metabolites (Fig. 4).

A possible explanation for the high degree of conservation of non-hub connectors is as follows. Connector nodes are responsible for inter-module fluxes. These modules are otherwise poorly connected or not connected at all to each other, so the elimination of connector metabolites will probably have a large impact on the global structure of fluxes in the network. On the contrary, the pathways in which provincial hubs are involved may be backed up within the module in such a way that elimination of these metabolites may have a comparatively smaller impact, which in addition would probably be confined to the module containing the provincial hub.

Our results therefore point to the need to consider each complex biological network as a whole, instead of focusing on local properties. In protein networks, for example, it has been reported that hubs are more essential than non-hubs³⁰. Notwithstanding the relevance of such a finding, our results suggest that the global role of nodes in the network might be a better indicator of their importance than degree²⁶.

Our ‘cartography’ provides a scale-specific method to process the information contained in the structure of complex networks, and to extract knowledge about the function performed by the network and its constituents. An open question is how to adapt current module-detection algorithms to networks with a hierarchical structure.

For metabolic networks—a comparatively well studied and well understood case—our method allows us to recover firmly established biological facts, and to uncover important new results, such as the significant conservation of non-hub connector metabolites. Similar results can be expected when our method is applied to other complex networks that are not as well studied as metabolic networks. Among those, protein interaction and gene regulation networks may be the most significant. □

Methods

Modularity

For a given partition of the nodes of a network into modules, the modularity M of this partition is^{11,18,21}:

$$M = \sum_{s=1}^{N_M} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right] \quad (1)$$

where N_M is the number of modules, L is the number of links in the network, l_s is the number of links between nodes in module s , and d_s is the sum of the degrees of the nodes in module s . The rationale for this definition of modularity is the following. A good partition of a network into modules must comprise many within-module links and as few as possible between-module links. However, if we just try to minimize the number of between-module links (or, equivalently, maximize the number of within-module links) the optimal partition consists of a single module and no between-module links. Equation (1) addresses this difficulty by imposing that $M = 0$ if nodes are placed at random into modules or if all nodes are in the same cluster^{11,18,21}.

The objective of a module identification algorithm is to find the partition with largest modularity, and several methods have been proposed to attain such a goal. Most of them rely on heuristic procedures and use M , or a similar measure, only to assess their performance. In contrast, we use simulated annealing²² to find the partition with the largest modularity.

Simulated annealing for module identification

Simulated annealing²² is a stochastic optimization technique that enables you to find ‘low-cost’ configuration without getting trapped in ‘high-cost’ local minima. This is achieved by introducing a computational temperature T . When T is high, the system can explore configurations of high cost whereas at low T the system only explores low-cost regions. By starting at high T and slowly decreasing T , the system descends gradually towards deep minima, eventually overcoming small cost barriers.

When identifying modules, the objective is to maximize the modularity, and thus the cost is $C = -M$, where M is the modularity as defined in equation (1). At each temperature, we perform a number of random updates and accept them with probability:

$$P = \begin{cases} 1 & \text{if } C_f \leq C_i \\ \exp\left(-\frac{C_f - C_i}{T}\right) & \text{if } C_f > C_i \end{cases} \quad (2)$$

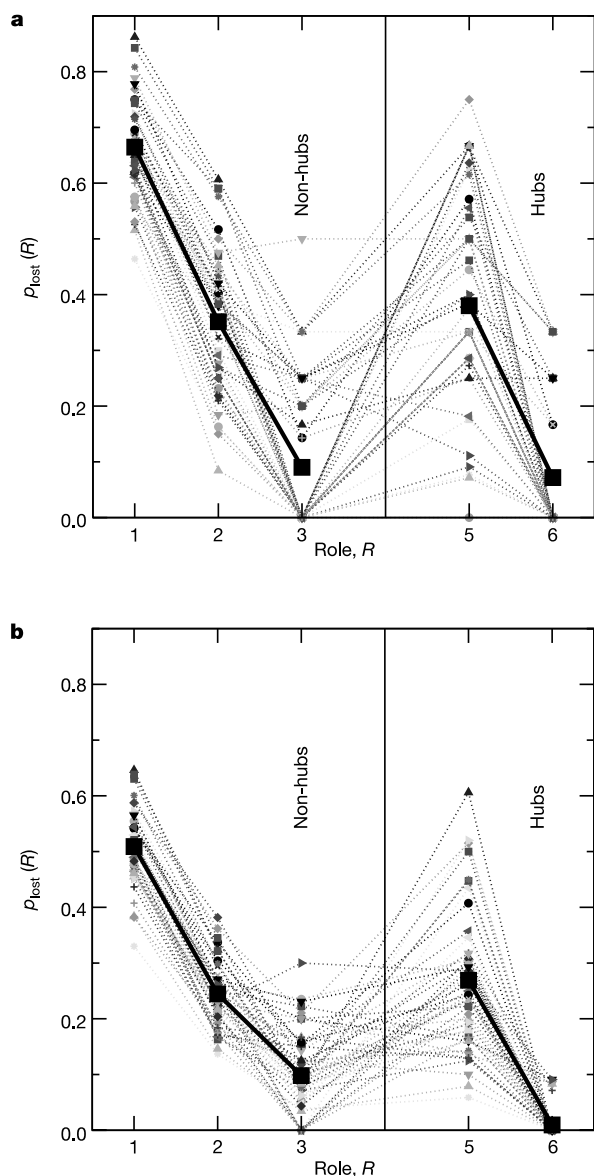


Figure 4 Roles of metabolites and inter-species conservation. To quantify the relation between roles and conservation, we calculate the loss rate $p_{\text{lost}}(R)$ of each metabolite (see Methods). Each thin line in the graph corresponds to a comparison between two species. Because we are interested in metabolites that are present in some species but missing in others, metabolic networks of species within the same superkingdom—bacteria, eukaryotes and archaea—are usually too similar to provide statistically sound information, especially for roles containing only a few metabolites. Therefore, we consider in our analysis only pairs of species that belong to different superkingdoms. The thick line is the average over all pairs of species. The loss rate $p_{\text{lost}}(R)$ is maximum for ultra-peripheral (R1) nodes and minimum for connector hubs (R6). Provincial hubs (R5) have a significantly and consistently higher $p_{\text{lost}}(R)$ than non-hub connectors (R3), even though the within-module degree and the total degree of provincial hubs is larger. Note that, out of the total 48 pair comparisons, only in two cases is $p_{\text{lost}}(R)$ lower for provincial hubs than for non-hub connectors, whereas the opposite is true in 44 cases. **a, b**, Results obtained for the MZ database (**a**) and the complete KEGG database (**b**).

where C_f is the cost after the update and C_i is the cost before the update.

Specifically, at each T we propose $n_i = fS^2$ individual node movements from one module to another, where S is the number of nodes in the network. We also propose $n_c = fS$ collective movements, which involve either merging two modules or splitting a module. For f we typically choose $f = 1$. After the movements are evaluated at a certain T , the system is cooled down to $T' = cT$, with $c = 0.995$.

Within-module degree and participation coefficient

Each module can be organized in very different ways, ranging from totally centralized—with one or a few nodes connected to all the others—to totally decentralized, with all nodes having similar connectivities. Nodes with similar roles are expected to have similar relative within-module connectivity. If κ_i is the number of links of node i to other nodes in its module s_i , $\bar{\kappa}_{s_i}$ is the average of κ over all the nodes in s_i , and $\sigma_{\kappa_{s_i}}$ is the standard deviation of κ in s_i , then:

$$z_i = \frac{\kappa_i - \bar{\kappa}_{s_i}}{\sigma_{\kappa_{s_i}}} \quad (3)$$

is the so-called z -score. The within-module degree z -score measures how well-connected node i is to other nodes in the module.

Different roles can also arise because of the connections of a node to modules other than its own. For example, two nodes with the same z -score will play different roles if one of them is connected to several nodes in other modules while the other is not. We define the participation coefficient P_i of node i as:

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{\kappa_{is}}{k_i} \right)^2 \quad (4)$$

where κ_{is} is the number of links of node i to nodes in module s , and k_i is the total degree of node i . The participation coefficient of a node is therefore close to 1 if its links are uniformly distributed among all the modules and 0 if all its links are within its own module.

Loss rate

To quantify the relation between roles and conservation, we calculate to what extent metabolites are conserved in the different species depending on the role they play. Specifically, for a pair of species, A and B , we define the loss rate as the probability $p(R_A = 0 | R_B = R) = p_{\text{lost}}(R)$ that a metabolite is not present in one of the species ($R_A = 0$) given that it plays role R in the other species ($R_B = R$). Structurally relevant roles are expected to have low values of $p_{\text{lost}}(R)$ and vice versa.

Received 17 August; accepted 16 December 2004; doi:10.1038/nature03288.

1. Amaral, L. A. N., Scala, A., Barthélémy, M. & Stanley, H. E. Classes of small-world networks. *Proc. Natl Acad. Sci. USA* **97**, 11149–11152 (2000).
2. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
3. Amaral, L. A. N. & Ottino, J. Complex networks: Augmenting the framework for the study of complex systems. *Eur. Phys. J. B* **38**, 147–162 (2004).
4. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular biology. *Nature* **402** (Suppl.), C47–C52 (1999).
5. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826 (2002).
6. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
7. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* **268**, 1803–1810 (2001).
8. Ma, H. & Zeng, A.-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277 (2003).
9. Hatzimanikatis, V., Li, C., Ionita, J. A. & Broadbelt, L. Metabolic networks: enzyme function and metabolite structure. *Curr. Opin. Struct. Biol.* **14**, 300–306 (2004).

10. Guimerà, R., Danon, L., Díaz-Guilera, A., Giral, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, no. 065103 (2003).
11. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, no. 026113 (2004).
12. Arenas, A., Danon, L., Díaz-Guilera, A., Gleiser, P. M. & Guimerà, R. Community analysis in social networks. *Eur. Phys. J. B* **38**, 373–380 (2004).
13. Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E. & Taylor, W. W. Compartments revealed in food-web structure. *Nature* **426**, 282–285 (2003).
14. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
15. Holme, P. & Huss, M. Subnetwork hierarchies of biochemical pathways. *Bioinformatics* **19**, 532–538 (2003).
16. Papin, J. A., Reed, J. L. & Palsson, B. O. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem. Sci.* **29**, 641–647 (2004).
17. Eriksen, K. A., Simonsen, L., Maslov, S. & Sneppen, K. Modularity and extreme edges of the Internet. *Phys. Rev. Lett.* **90**, no. 148701 (2003).
18. Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69**, no. 066133 (2004).
19. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. Natl Acad. Sci. USA* **101**, 2658–2663 (2004).
20. Donetti, L. & Muñoz, M. A. Detecting network communities: A new systematic and efficient algorithm. *J. Stat. Mech. Theor. Exp.*, P10012 (2004).
21. Guimerà, R., Sales-Pardo, M. & Amaral, L. A. N. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E* **70**, no. 025101 (2004).
22. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
23. Wasserman, S. & Faust, K. *Social Network Analysis* Ch. 12, 4 (Cambridge Univ. Press, Cambridge, 1994).
24. Guimerà, R. & Amaral, L. A. N. Cartography of complex networks: Modules and universal roles. *J. Stat. Mech. Theor. Exp.* P02001 (2005).
25. Rives, A. W. & Galitski, T. Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA* **100**, 1128–1133 (2003).
26. Han, J.-D. J. *et al.* Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**, 88–93 (2004).
27. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
28. Schuster, S., Fell, D. A. & Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.* **18**, 326–332 (2000).
29. Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. & Dandekar, T. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Micropasma pneumoniae*. *Bioinformatics* **18**, 351–361 (2002).
30. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Broadbelt, V. Hatzimanikatis, A. A. Moreira, E. T. Papoutsakis, M. Sales-Pardo and D. B. Stouffer for discussions and suggestions, and H. Ma and A. P. Zeng for providing us with their metabolic networks' database. R.G. thanks the Fulbright Program and the Spanish Ministry of Education, Culture & Sports. L.A.N.A. acknowledges the support of a Searle Leadership Fund Award and of a NIH/NIGMS K-25 award.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.A.N.A. (amaral@northwestern.edu).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

European Sales Manager:

Andy Douglas (4975)

UK/ RoW/ Ireland:

Nils Moeller (4953)
Irene Vigila-Alton (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Advertising Production

Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Niamh Shields

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan

Tel + 49 89 54 90 57 11

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harakakatamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

Job movements

From mergers and acquisitions to a move towards outsourcing, the pharmaceutical employment landscape is being tugged by global forces, sending jobs in different directions. Development and manufacturing jobs are shifting from the United States and Europe to overseas (see page 902). Meanwhile, mergers and cost-cutting could reduce the number of overseas sales jobs. Analysts speculate that Pfizer, which, after a merger with Pharmacia, is now the world's largest drug company, is set to shed up to 12,000 jobs from its global workforce of 122,000. The question is, which jobs will be cut, and how will the rest of the industry respond?

The conventional wisdom is that most of the cuts will come from Pfizer's global sales force — and most of those from outside the United States. But any cuts to the sales force will have broader repercussions. Inside the company, employees are apt to feel less secure about their future. And although the company is building a US\$35-million research facility in New Haven, Connecticut, concerns about long-term stability might mean it will have a hard time filling it with scientists from outside the firm.

Analysts predict that other companies will follow Pfizer's lead and trim their global sales forces — and perhaps other employees, too. This merely illustrates a trend that industry-watchers have been plugging for years: the drug industry is moving away from drug development and sales towards discovery and marketing.

But jobs in drug development won't disappear completely. Discovery will still occur at some level, and will continue to expand in mid-sized and large biotech companies. Development and manufacturing will tend to be outsourced to India and China, as those countries continue to increase their capacity and offer lower labour costs. At this point, it is only the sales force that will be asked to do more with less.

Paul Smaglik

Naturejobs editor



Contents

CAREERS AND RECRUITMENT

Global outsourcing in
the drug industry p902

CAREER VIEW

Scientists & Societies
Horizons in Molecular Biology
Graduate Journal
A story to tell
Movers
Arne Henden p904

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

India in demand

What's a company to do when it needs faster, cheaper new drugs and chemists are hard to find? Look for a source of bright graduates with low living costs, where legal changes have pushed firms to seek work, and you're there, says Emma Marris.

The pharmaceutical industry is limping a bit, especially in the West. Some of its biggest-selling drugs have been brought down or impugned by scandal — painkillers linked to heart attacks, antidepressants blamed for increasing the risk of suicide — and new drugs are oozing ever more slowly out of the pipeline. In response, many companies want to speed up drug discovery yet cut costs at the same time.

The way to meet these seemingly mutually exclusive goals may lie through global outsourcing. Other industries, such as finance and manufacturing, have been hiring workers in foreign countries for years, saving money on wages and benefits. In Britain and the United States, consumers have come to expect their customer-service calls to be answered by staff in India.

Some drug companies have relocated administrative tasks, including payroll and computer support, overseas. A report from management consultants A. T. Kearney, published at the end of last year, said that the top ten drug companies could save US\$8 billion in ten years of outsourcing “general and administrative” tasks globally. Drug manufacturing, too, is increasingly being moved abroad.

A substantial number of companies are now also outsourcing hard science. Although early drug discovery is still mostly domestic, everything in the pipeline after that — from early development to clinical trials — can be farmed out. Companies are understandably cagey about discussing what parts of their operations they are outsourcing, and even what to call their far-reaching international workforce strategies (see ‘Other arrangements’, opposite), but the trend is clear to industry observers.

“We are in the early part of the wave as far as outsourcing to India and China is concerned,” says Nailesh Bhatt, chief executive of Proximare, a consulting company in Franklin Park, New Jersey, that helps drug companies to outsource. The impetus for technical outsourcing is a combination of cost-cutting and scientific supply and demand.



Growth industry: GVK Biosciences in Hyderabad plans to double its staff.

GVK BIOSCIENCES

Chemistry provides a good illustration of India's attractions, in both economic and scientific terms. Bhatt estimates that the salary, overheads and benefits for one freshly minted American PhD come to an annual figure of \$230,000. An Indian scientist would cost about a third of that.

BREAKING THE BOTTLENECK

But the issue isn't just cost, says Sam Tetlow, a principal at Research Triangle Ventures, a North Carolina venture-capital firm. Tetlow interviewed 72 drug and biotech companies in the United States, Europe and Japan to find out about their outsourcing plans and experiences. “There are hardly any chemists in the world, and that's where the bottleneck is,” he says. Meanwhile, 122,000 chemists and chemical engineers graduate in India each year, according to the Kearney report. In the West, there has been a consistent shortage of medicinal chemists over the past several years (see *Nature* **424**, 594–596; 2003).

India is especially suited to fill this gap, thanks to international trade law. For years, Indian chemists were kept in demand by a lucrative trade in reverse-engineering popular drugs to make generics, says David Templeton, who represents the Indian company SIRO Clinpharm in the United States. That changed on 1 January, when India fully joined the Trade-Related Aspects of Intellectual Property Rights agreement (TRIPS). This closed the legal loophole that allowed India to make generic versions of drugs still under Western patent protection. Now many of the Indian drug firms that used to make generics are turning to contract research and manufacturing.

Many contract research organizations (CROs) work on the process of figuring out, step by step, what a

Biotech outsourcing stays domestic

Big drug companies are not the only outsourcers, according to Sam Tetlow of venture-capital firm Research Triangle Ventures. Biotech companies are nearly as enthusiastic, although they are not going abroad so much. Biotechs are usually small and don't have the staff to do everything themselves. They also work notoriously close to the bone, and the cost savings to be made offshore are tempting.

Tetlow points out that partnering with established companies can increase the credibility of a scrappy start-up. In a survey he did of biotech companies around the world, he found that they were more likely than the big drug firms to list “utilization of external expertise” as their top reason for outsourcing.

E.M.



D. H. WELLS/CORBIS **First in line: India is increasingly the destination of choice for drug companies seeking to outsource elements of their operations.**

drug does in a person's body in order to prove both that it is safe, and that it has some effect. Half of the companies that Tetlow talked to plan to outsource at least 80% of this work by 2008.

Another hot area is animal-model pharmacokinetics, using animal models of human disease to see how drugs metabolize within the body before they are tested in humans. Also popular is lead optimization, putting promising compounds through their paces to see if they are worth developing. These two processes weed out drugs that looked promising in screens or in animals but don't work in humans, preventing costly human trials that are doomed from the beginning.

Sanjay Reddy, chief executive of Hyderabad-based GVK Biosciences, stresses that the next challenge for Indian CROs is to stop promoting themselves in terms of cost savings and start stressing their value-adding expertise. His company serves seven of the biggest ten

drug companies, he says, and specializes in medicinal chemistry, cheminformatics, bioinformatics and clinical research. He hopes to double the number of scientists working for him in the next year, and expand to cover discovery, preclinical, clinical and chemical development. "It's a very exciting time," he says.

At Dr Reddy's Laboratories in Hyderabad, Anji Reddy is prepared to adjust. "From generics to new drugs, from imitation to innovation, is a big leap but there is growing optimism that Indian companies have an opportunity in the R&D space," he says.

GROWING PAINS

There are, of course, disadvantages to outsourcing preclinical work abroad. Doing business half-a-world away is not easy. Keeping in verbal touch with your partners means waking up early or going to bed late.

As in the beginning of all things, there is an ad hoc feel to some of these arrangements. As time goes by, drug companies will formalize their strategies, Bhatt guesses, and CROs will be more stable. "These companies are just getting started," he says.

Chris White of A. T. Kearney agrees that the industry is still exploring outsourcing options. "They are really just trying to show proof-of-concept and work out the bugs," he says. But along with other observers, he feels that an increase in outsourcing is inevitable. "Those that don't follow the pack will be left behind," he says.

According to his survey of drug companies, Tetlow says, the number-one reason for outsourcing in general is "focusing on core competency". In other words, it makes sense for these companies to stick with the things they do well, and hire out the rest.

This suggests a natural limit to outsourcing. There is some work that is so essential to what made a company successful in the first place that it would make no sense to hire someone else to do it. And there might be some projects so classified that companies would want to keep them close to the chest. ■

Emma Marris is a freelance science writer based in Washington DC.

Other arrangements

Drug companies are opening facilities in India and establishing different models of collaboration.

AstraZeneca recently opened a research centre in Bangalore specializing in tuberculosis. Carol Cruickshank of management consultants A. T. Kearney was duly impressed when she visited the centre during a trip to India to look at the outsourcing picture. "If you were dropped by helicopter into that facility, you'd have no idea you were in India," she says. "It's nicer than AstraZeneca's facility in Wilmington."

Since October 2003, GlaxoSmithKline has had an agreement with Ranbaxy Laboratories in New Delhi to do drug-discovery work. For GlaxoSmithKline, it is one of a number of international collaborations that set up what spokesman Rick Koenig calls an "alternative drug-discovery initiative" to complement its in-house efforts.

E.M.

GRADUATE JOURNAL

A story to tell

About three months ago, I applied to give a talk at a major meeting. I wasn't sure what my chances were, but I think my research story is interesting and covers an area that is not that well known (pairing-dependent gene expression). Last week I found out that I have been accepted — and now I have some mixed feelings about doing the presentation.

On the positive side, the talk is short but the meeting is big. Even though I won't be continuing in this field, I am excited to share the work I have done. But giving a short talk at a conference presents some unique challenges. To be successful, I will need to be clear yet concise — even more difficult in a limited time frame. This seems especially hard given the diversity of the audience as well as the relative obscurity of my field.

I accepted the challenge because the positives outweigh the negatives. I've given a great deal of thought to what I hope to accomplish at this meeting. I will have the chance to share my story, one that I have been working on for the past four years. And I will gain the experience of talking in front of an audience from outside my institution. I've heard from several different sources that this is both a good experience and good for the CV. Now it's time to prepare and practise. ■

Anne Margaret Lee is a graduate student at Harvard University.

SCIENTISTS & SOCIETIES

Horizons in Molecular Biology

Most scientific conferences are organized by senior scientists for senior scientists. They often cater only for a specialized audience with an advanced scientific background. As a result, few young researchers have the opportunity to be involved in planning these events or in presenting at them.

In response, a group of molecular-biology PhD students at the International Max Planck Research School in Göttingen, Germany, decided in 2003 to organize a symposium by young researchers for young researchers. Rather than focus on a specialized audience, we sought to attract early-stage researchers with an open outlook on alternative and novel scientific approaches and their related technologies.

The resulting conference series, Horizons in Molecular Biology, aims to equip young researchers with a comprehensive overview of current frontier research in the life sciences. But learning about the cutting edge is only one goal of these meetings.

We want to help establish an interdisciplinary scientific network among today's PhD students. The Horizons meetings place a strong emphasis on networking. Various social and scientific events catalyse information exchange in an informal atmosphere.

The Horizons meetings have already benefited the young scientific community in Göttingen. Those involved in organizing the symposia have been presented with a unique opportunity to gain experience in fundraising, design, promotion and event management. The first symposium allowed

the organizing committee to forge invaluable contacts, improve team-building and communication skills, and gain new insights into the current scientific landscape.

The second Horizons symposium — Decoding Nature: Hierarchy of Interactions — will take place on 17–19 March at the Max Planck Institute for Biophysical Chemistry in Göttingen. The programme will include a session of talks given by PhD students and a large forum for poster presentations. The social programme features, among other events, a party with live music. As a result, students who attend will feel like they are among their peers, rather than on the outside trying to get in with their more established colleagues. ■

Ralf Jauch, a Horizons organizer, is a graduate student at the Max Planck Institute for Biophysical Chemistry in Göttingen, Germany.
♦ www.horizons.uni-goettingen.de

MOVERS Arne Henden, director, American Association of Variable Star Observers, Massachusetts



Arne Henden's first exposure to the wonders of star-gazing came as a child growing up in Arizona, when his father took him to Lowell Observatory in Flagstaff. After the junior astronomer peered through the 24-inch Alvin Clarke telescope, which Henden describes as "one of the long tube ones with all the knobs at the end", he was hooked.

The experience stayed with Henden, and he pursued a formal training in astronomy. He learned early on that there are relatively few faculty positions in the field, so he aimed for contract work, where opportunities are more abundant. He got his first contract

research job while still a graduate student.

"Eventually I worked my way back to Flagstaff," says Henden. There, at the US Naval Observatory, he explored his interests in optical and near-infrared imaging, variable stars and γ -ray burst afterglows. And he also began mentoring amateur astronomers.

In 1997, Henden became involved with the American Association of Variable Star Observers (AAVSO) in Cambridge, Massachusetts, when he provided photometry for variable-star charts. Soon afterwards, he became chief adviser to the AAVSO International High Energy Network.

Henden's new position as director of the AAVSO will allow him to concentrate more on that interest. The AAVSO has a membership of 1,200 amateur astronomers, and Henden is excited about mentoring this large pool of talent. "They are all very enthusiastic, so it's fun

to work with them," he says. His new job will help the astronomers improve their skills through tutorials and workshops, and Henden will oversee the collation of their results into the association's database.

But there are trade-offs. The AAVSO doesn't have its own telescope, so Henden's own observing time will be limited. He'll have a smaller scientific staff, with only one postdoc to analyse data produced by its members. And he'll no longer be involved in developing new technology. Instead he'll be focused on running the organization. "That comes first," Henden says. "My own science comes second."

Directing the AAVSO is a fitting career step for someone who got his first taste of the cosmos as an amateur astronomer. Henden's position will allow him to help others move, as he did, from awe and wonder to contributing to a real understanding of the night sky. ■

CV **1993–2005:** Senior research scientist, US Naval Observatory, Flagstaff, Arizona.
1985–93: Research associate, Ohio State University, Columbus, Ohio.
1979–85: Research associate, Goddard Space Flight Center, Greenbelt, Maryland.

The party's over

It was only a game....

Penelope Kim Crowther

"Tell me again about the balloons, Grandpa."

"Now? We're nearly there."

"Pleaseee."

"Okay, okay. Well, when I was a kid we used to have balloons at our birthday parties that were filled with ... helium."

"No!!!"

"Yes. And when we were done playing with them, we'd sometimes let them just drift off into the sky. Or swallow the helium to make our voices go squeaky."

"You ATE it?"

"You bet. But it was cheap back then — not like now. It wasn't till your Grandma and me were about your age that people started to realize how much helium we needed, and how we were running out. Physicists were doing more and more clever things at very cold temperatures, and they needed a lot of helium to keep things cold. But there just wasn't that much about. Do you remember where helium comes from?"

"Uh ... no."

"Sophie ... I've told you a hundred times. Hang on a minute, I just have to pay this road toll. There. So, all our helium is left over from when the planet first formed. It leaks out of the middle of the Earth and trickles out of the ground, and then it hangs about in our air for a while. But it's so light that eventually it spins out of the sky into outer space. Our planet leaks helium — like a pricked balloon."

"So if it just comes out of the ground, why did we run out?"

"Well it's hard to collect something when there's so little of it in the whole sky. Sucking helium out of the air is like mining a beach for lost wedding rings. Not a very good idea. The only place where there's enough of it to mine is ... in oil. Oil collects helium and stores it up."

"But there isn't any oil now."

"Exactly. Back in the '20s, we were running out of oil pretty quickly and all the helium plants in Texas and Saudi Arabia were running down as fast as the oil wells were running dry. So that's why your Grandma, who was a very clever geologist, made so much money. She was young and adventur-



ous — like you." Sophie giggled at that. "And she saw what was happening, so she went out to Nepal, where the Earth is all cracked up from earthquakes and there are places where helium leaks out in bigger streams than anywhere else on the planet. And the American company she worked for — still works for — built big buildings over these cracks so they could collect the helium. And then they sold it on."

"I'm clever too, you know."

"I know you are. Here we go — here's our turning."

"So is that why Granny still lives in the United States?"

"In a way. Your Grandma decided that all the money from that helium was more important to her than other things ... which is why you get to have such nice birthday presents, even if you don't have helium-filled balloons. And that's why I moved here ...

to do other things instead. Now, we're here!"

Sophie looked up and saw the sign: 'ITER — the future of fusion'. She knew this place, it was where her Grandpa worked sometimes. It had a sign outside with a picture of a little explosion. Sunny rays were flying out, which was energy you could use to make your computer work, she knew. Grandpa always said this would be the next big thing, after oil. The picture showed other things flying out too, little circles labelled 'D' and 'n' and 'He'. "He who?" she wondered.

Grandpa pulled into the gravel drive and stopped the car. "Now you'll stay in the car while Grandpa has his meeting? I'll only be a little while."

Sophie waited until she saw him go inside the gate, and then the doors. Then she got the birthday present out of her bag and gave it a little shake. Grandma had given it to her the week before, and told her it was a surprise for the man who ran ITER. She was supposed to hide it, that was the game, while Grandpa was at his meeting, but not to tell anyone. Sophie liked surprises. Beneath the wrapping and the bows the present ticked like a clock. She got out and hid it in the bushes by the gates, and went back to wait.

They didn't feel the explosion ten hours later, when they were safely back in the flat having tea. Not until Grandpa turned on the news the next morning did Sophie see the building on fire, big billows of black smoke heading for the sky. She could hear the reporter over the sound of sirens: "... This will put back fusion research for years. On the eve of a demonstration project that was meant to prove to the world the reality of cheaper, cleaner energy, there is only smoke to show for decades of..."

Sophie didn't understand what was going on. Or why Grandpa had his head in his hands. Was he crying? The newscast went on: "In related news, the price of helium jumped sky high today as the potential source from fusion plants dried up — for decades now at least..."

Penelope Kim Crowther is a journalist and occasional writer of fiction, who tries very hard not to muddle up her genres.